THE UNIVERSITY OF CHICAGO

PROCESSING CONTEXT-SENSITIVE EXPRESSIONS: THE CASE OF GRADABLE ADJECTIVES AND NUMERALS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE HUMANITIES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF LINGUISTICS

BY
HELENA APARICIO TERRASA

CHICAGO, ILLINOIS
MARCH 2018

To my family and especially to its two most recent additions, Fer and Pedro.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am grateful for having had the chance of spending the past six years doing research in the Linguistics Department at the University of Chicago. Working towards a PhD can be as exhilarating as it can be disorienting. It is especially due to the latter that I would like to express my sincere gratitude to my two advisors, Ming Xiang and Christopher Kennedy, for all the guidance they have provided me over the past years. Ming's intellectual sharpness and rigor, and Chris' insightfulness and ability to always see the path forward have been and continue to be qualities that I strive for. Thanks also to Gregory Kobele, Karlos Arregi and Itamar Francez from whom I have learned a lot while in Chicago, and to Marcel den Dikken and Chris Barker who taught me more than I was probably able to absorb during the years in New York.

The graduate students and postdocs with whom I have overlapped during these years have also been a very important part of this experience. I would like to thank Julian Grove for all the good, linguistics and otherwise, conversations and the even better laughs. *Gracie* to Andrea Beltrama and Livia Garofalo, for their friendship and for being so caring. A very special thanks goes to Katie Franich, whose friendship means more to me than she probably knows! Thanks to Christina Kim for introducing me to the world of the Visual World paradigm and to Timothy Leffel for all the discussions about adjectives and meaning. A big thanks to Kate Christensen for being a fantastic Research Assistant over the summer of 2016, and to all the members of the Language Processing Lab for all the feedback they have provided over the years to different states of the work contained in this dissertation. Thank you to all my cohort members, Tamara Vardomskaya, Asia Pietraszco, Gallagher Flinn, Katie Franich, Diane Rak and Mike Pham, with whom I got to spend a lot of time solving problem sets during the first two years of grad school. I have very fond memories of those times! Thanks also to Ryan Bochnak, Peet Klecha, Jackson Lee, Martina Martinović (and Ben!), Emily Hanink, Jeff Geiger, Josh Falk and Adam Singerman.

Outside the Department, there have also been people, both in Chicago and back home, whose friendship has shaped the past years. *Mil gràcies* to Mar Rosàs, Miquel Saumell and Samuel Fox

for bringing with them a piece of home to Hyde Park. *¡Gracias chavalada!* for all the dinners, bbq's, *coves*, *yimmies*, *maravillas*, *San Givings* and old fashions. *Gràcies* to Maria del Mar Vanrell who made my arrival to Boston after grad school a wellcoming experience. *Gràcies* to Cristina Aliagas for offering me the kind of friendship that only gets better with every passing year, and to Anna Serra with whom I share every single day despite having lived in different continents for almost ten years now. *Gràcies per tot, Anna*.

Mamà, papá, Biel i Ita: *Gràcies per la paciència, pels ànims constants i per fer-me sentir sempre tan acompanyada tot i la distància. Vos estim.*

Finally, thanks to my husband, Pedro, whose support, determination, and love motivate me, challenge me and uplift me every step of the way.

ABSTRACT

This dissertation investigates the processing of two types of meaning and context interactions, vagueness and imprecision, through the case study of gradable adjectives and round numerals. The first half of the dissertation, asks the question of whether vagueness and imprecision should be collapsed into one single category, or whether they should be treated as fundamentally different types of meaning and context interactions. I investigate this question through two experiments (a Visual World eye-tracking study and a judgment task study) that focus on the processing of Relative and Absolute adjectives. Existing accounts of the relative vs. absolute distinction agree in that the context-sensitivity displayed by Relative adjectives is due to vagueness. More specifically, vagueness results from the fact that these adjectives have highly flexible lexical thresholds, whose value is generally fixed by accessing contextual information. There is however less consensus about whether context-sensitive interpretations of Absolute adjectives result from threshold variability in the sense described above, or from pragmatic reasoning about imprecision. The results reported in the dissertation converge to show that participants recruit and integrate information from the visual context differently during the processing of Relative and Absolute adjectives, suggesting that the context-sensitivity of these two classes of adjectives is indeed of a different nature. I argue that these findings constitute support for theories that claim that Absolute adjectives are not lexically context-sensitive—and therefore have fixed, context-insensitive, adjectival thresholds—, and that variable interpretations of Absolute adjectives involve imprecision.

In the second half of the dissertation, I investigate the processing of imprecision in more detail. It has been claimed that comprehenders favor imprecise interpretations over precise ones whenever possible. One of the explanations that has been put forth to explain this alleged preference is that imprecise representations might be less costly to process for the comprehender. The few existing studies that have sought to empirically substantiate this claim focus on the numeral domain, and use the round vs. non-round distinction (i.e. 100 vs. 101) as a proxy for imprecise vs. precise interpretations of the numerals. The logic behind this choice is that non-round numbers usually

give rise to precise interpretations, while round numbers (by assumption) tend to be interpreted imprecisely. I argue that approaching this question from this perspective introduces the confound that non-round numerals might be independently difficult to process for reasons that are orthogonal to (im)precision calculation. I suggest that the relevant comparison should therefore be between precise and imprecise interpretations of round numbers. With these goals in mind, I conducted a series of three self-paced reading studies, where I tested (im)precise interpretations of the same round numbers. Contra previous claims, the results show that imprecise interpretations are not faster to process than their precise counterparts, but rather the opposite: imprecise interpretations incur a processing penalty compared to precise interpretations, independently of whether (im)precision is signaled explicitly by means of a slack regulator (e.g. *about* or *exactly*), or through pragmatic cues (e.g. reasoning about conversational goals).

CHAPTER 1

INTRODUCTION

1.1 Context-sensitivity in natural language

More often than not, assigning an interpretation to a sentence requires understanding what its conventional meaning *means in context*. Conversational implicatures (1) are a hallmark of this interaction between meaning and context. In (1), nothing about the conventional meaning of B's answer can possibly lead to the inference in (1c).

(1) a. A: Do you speak Portuguese?

b. B: *My husband does*.

c. Implicature: B does not speak Portuguese.

Rather, the implicature in (1c) is triggered by flouting the Gricean maxim of relevance, since—strictly speaking—(1b) is irrelevant as an answer to A's question. Conversational implicatures like (1c) are highly context-dependent, as shown by the fact that not any flouting of the maxim of relevance would warrant the implicature in (1c). If B's answer had been *My neighbor's barber does*, A would have been more puzzled than compelled to infer (1c), unless she was aware of some salient relationship between B and his neighbor's barber. It is precisely the world knowledge about the privileged status that the relationship between husband and wife has in our culture coupled with the flouting of the maxim of relevance what leads A to infer (1c).

The interaction between meaning and context exemplified in (1) lives in the discursive dimension. However, beyond this discursive type of meaning and context interactions, natural languages are filled with instances where the interplay between meaning and context is actually grounded at the world level. To put it differently, the meaning of certain grammatical categories is such that its interpretation changes as a function of context. Consider examples (2)-(4).

(2) John is *tall*.

(3) *1,000* fans attended the concert last night.

(4) *All* passengers were at gate 13.

Each of the examples in (2)-(4) represents a possible way in which semantic meaning can interact with context resulting in a new interpretation. For instance, there exist expressions that are intrinsically context-sensitive. This is the case of the relative adjective *tall* in (2), whose interpretation varies depending on what kind of individuals John's height is being compared to (e.g. male adults, toddlers, basketball players, etc.). Without information about this implicit class of individuals compared to which John's hight counts as tall—in the speaker's opinion—, it is not possible for the comprehender to infer what range of degrees of height would make sentence (2) true.

Example (3) represents another type of interaction between meaning and context. In its imprecise interpretation, the round number in (3) can be construed as a fuzzy interval rather than a precise measure (e.g. an interval with fuzzy boundaries around 900 and 1,100). In such a case, the context-sensitive interpretation causes the semantic truth-conditions of the numeral to loosen, since the amount of values that the numeral can take under this new reading is a superset of its original semantic denotation. In other words, the context-sensitive interpretation is asymmetrically entailed by the denotation of the numeral (Lauer 2012).

The opposite type of meaning and context interaction is also attested (4). The strictest interpretation of sentence (4), where the restrictor of the universal quantifier is the set of all passengers in the world of evaluation, is most probably guaranteed to always be false. In order to make a sentence like (4) true and usable at all, the truth-conditions of quantifiers like *all* must be checked against some contextually salient subset of individuals, much like we discussed for the case of relative adjectives, that restricts the domain of the universal quantifier. In the case of (4), this would be the set of passengers of the flight leaving from gate 13. Factoring in this contextual informational has the effect of strengthening the interpretation of the expression. In this respect, the interaction between meaning and context displayed by (4) can be seen as the mirror image of the context sensitive interpretation of (3).

Despite these surface differences, the examples in (2)-(4) can also be seen as displaying a single type of sensitivity to context. A type of context-sensitivity governed by a heuristic bias dictating that information that is retrievable from context (linguistic or otherwise) should not be linguistically encoded. Such logic can actually be traced back to the Gricean maxim of quantity, and has been used to account for the existence of ambiguity in natural language (Piantadosi *et al.* 2012). After all, the amount of context-sensitivity displayed by (2)-(4) could have been greatly reduced exclusively by linguistic means, as exemplified in (5)-(7).

(5) John is 7 feet *tall*.

(6) Exactly *1,000* fans attended the concert last night.

(7) *All* the passengers of flight AA416 were at gate 13.

An important question is whether all the cases of context-sensitivity exemplified in (2)-(4) can be explained as a result of this trade-off between linguistic and other sources of contextual information, or whether the differences discussed with respect to these examples uncover deeper properties about the grammar and pragmatics of these lexical items (or a combination of both, of course). Therefore, an important question is whether context-sensitivity should be encoded in the grammatical representation of these lexical items, or whether some or all of the discussed context-sensitive interpretations are the result of pragmatic reasoning alone, and if so what exactly triggers these pragmatic inferences and how these are modulated by the semantic representation of the lexical item itself. Settling these questions on purely theoretical and intuition-based grounds has proven a difficult task, as will be discussed in Chapter 2. However, developments in the field of language processing in the past decades have equipped us with new tools to address some of these issues. In the remainder of this chapter, I discuss some of the breakthroughs made by psycholinguists regarding the question of how humans capitalize on contextual information to decode the linguistic signal as it unfolds in real time, and point at some of the ways in which these discoveries can help us approach the questions addressed in this dissertation.

1.2 Insights from online language processing

An alternative route to approach this question is to turn to online processing. Over the last decades, research in psycholinguistics has amassed a convincing amount of evidence showing that linguistic interpretation is highly incremental (Altmann & Steedman 1988; Tanenhaus *et al.* 1995; Sedivy *et al.* 1999; Chambers *et al.* 2002; Kamide *et al.* 2003, among many others), and that humans quickly integrate information from a variety of sources as the linguistic input unfolds to guide their expectations about plausible continuations. Such sources of information relate to different levels of linguistic representation such as syntactic structure (Gibson 1998, 2000; Hale 2003; Levy 2008; Lewis & Vasishth 2005) or lexical-semantic information (Trueswell *et al.* 1994; Wolter *et al.* 2011; Kim *et al.* 2015). More relevant for the current dissertation are findings showing that humans are highly sensitive, from the very early stages of processing, to a wide variety of information about the linguistic and non-linguistic context. These sources of information range from low level implicit knowledge about word frequencies to higher level aspects of communication such as the identity features and idiosyncrasies of the interlocutor, or the make up of the surrounding visual context (Rayner 1998; Sedivy *et al.* 1999; Kamide *et al.* 2003; Grodner & Sedivy 2011; Pogue *et al.* 2016). The methodologies used by some of these studies provide us with ways to quantify the influence of context during online language processing with millisecond accuracy. Therefore, the time is ripe to capitalize on these methods and use language processing data to further inform the existing theories of those lexical items that display context-sensitive uses.

1.3 Questions addressed in this dissertation

The first half of the dissertation focuses on question (8):

(8) Should vagueness and (im)precision be collapsed into one single type of context-sensitivity?

In order to answer question (8), I look at Gradable Adjectives, a class of adjectives formed by Relative and Absolute Adjectives, both of which present context-sensitive uses. Some theories of

the relative vs. absolute distinction treat the context-sensitivity of Relative and Absolute adjectives as fundamentally different: Relative Adjectives give rise to vagueness, whereas Absolute adjectives give rise to imprecision, which unlike vagueness is argued to result from pragmatic reasoning. Other theories treat the context-sensitivity displayed by these two types of adjectives as a unified phenomenon, thus collapsing vagueness and imprecision into one single class. This dissertation provides behavioral evidence supporting the former view.

The second half of the dissertation looks in more detail at the processing of imprecision through the case study of round numbers. In particular, I address the question in (9).

(9) Are imprecise representations of number words costlier to process for the comprehender?

I tackle this question through a series of self-paced reading studies, where I compare the processing of imprecise interpretations of round numbers against the processing of their precise counterparts. Contra previous claims, the results reported in this dissertation suggest that imprecision is costlier than precision.

1.4 Outline of the dissertation

Chapter 2 lays out the theoretical background for Experiments 1 and 2. I start by providing an overview of the notion of *gradability* in the adjectival domain, and spell out the grammatical properties of gradable adjectives, namely Relative and Absolute adjectives. I then move on to characterize the different types of context-sensitivity shown by these two classes of adjectives—which will constitute the focus of the sequence of experiments presented in Chapter 3—, and introduce the existing theories tailored to account for the relative vs. absolute distinction. I conclude the chapter by reviewing previous Visual World eye-tracking studies about online processing of prenominal adjectives, and argue that this experimental paradigm is particularly well-suited to address the research question in (8).

Chapter 3 presents the results of Experiments 1 and 2. Experiment 1 consists of a VW eye-

tracking study investigating whether information from the visual context is integrated in comparable ways during online processing of relative and absolute adjectives. Results show that the processing signature of relative and absolute adjective is indeed qualitatively very different, lending support to the view that these two classes of adjectives engage with context in fundamentally different ways. Taken together, these findings are best accommodated by theories that assume that relative adjectives give rise to lexically encoded vagueness, whereas absolute adjectives give rise to pragmatic reasoning about imprecision. I therefore conclude that vagueness and imprecision are two types of context-sensitivity that do not form a natural class, despite surface appearances.

Experiment 2 is conceived as a follow up to Experiment 1. In it, I explore the strength of pragmatic contrastive inferences associated with the adjective classes tested in Experiment 1 when used prenominally, a syntactic position that often correlates with a restrictive interpretations of the adjectival predicate. The goal of the experiment is to ensure that the differences observed between relative and absolute adjectives in Experiment 1 are not the result of general quantity and manner based pragmatic reasoning. Results suggest that, while there exists a deep connection between reasoning about informativity and the patterns of effects obtained in Experiment 1, quantity and manner based pragmatic reasoning alone cannot account for the different processing signatures displayed by relative and absolute adjectives, thus reinforcing the conclusions of Experiment 1.

In the second half of the dissertation, I change gears and focus more deeply on the processing of imprecision. I do so by looking at the case study of round numerals. **Chapter 4** provides an overview of the characteristics exhibited by variable uses of round numbers, which, in their context-sensitive interpretation, behave like Absolute Adjectives. I then proceed to discuss the notion of *cost* as a driver of the preference of approximation, a heuristics that states that comprehenders should default to the most imprecise interpretation licensed by the context. I conclude the chapter with a discussion of the few studies that have investigated in one way or another the processing of imprecision.

Chapter 5 contains results from three self-paced reading studies (Experiments 3, 4 and 6) and a

judgement task (Experiment 5) that investigate in two situations: when the imprecise interpretation induced by a Slack Regulator such as *approximately* (Experiments 3 and 4), or cases in which the comprehender arrives at the imprecise interpretation as a result of pragmatic inferences about conversational goals (Experiment 6). Collectively, the results of these experiments show no processing advantage advantage for imprecise interpretations. Rather, the opposite is found: imprecise interpretations take significantly longer to process than their precise counterparts. Finally, in **Chapter 6**, I conclude and point at future questions that the current results bring out.

CHAPTER 2

CONTEXT-SENSITIVE INTERPRETATIONS OF GRADABLE ADJECTIVES: THE RELATIVE VS. ABSOLUTE DISTINCTION

2.1 Introduction

Both relative adjectives like *big* and absolute adjectives like *empty* are sensitive to the non-linguistic context: in the former case, the context determines how much size is required to count as big; in the latter, the context determines how much deviation from total emptiness is allowed to count as empty. Whereas it is generally agreed that the role of context with relative adjectives is to fix the value of a threshold variable, the status of absolute adjective thresholds, and therefore the role of context in their interpretation, remains an object of debate. Some researchers have argued that all gradable adjectives have context-sensitive threshold variables that are assigned values by the same mechanisms (Lassiter & Goodman, 2013). Others have claimed that absolute adjectives have fixed, endpoint-oriented meanings and that sensitivity to context arises from pragmatic reasoning about imprecision (Kennedy 2007; Syrett *et al.* 2009; van Rooij 2011; Burnett 2014; Qing & Franke 2014; Leffel *et al.* 2016).

In this chapter, I review theoretical accounts of the relative vs. absolute distinction within the class of gradable adjectives, and the role that context plays in the interpretation of these predicates. The chapter proceeds as follows. Section 2.2 summarizes the different grammatical properties of relative vs. absolute adjectives with special emphasis in describing the ways in which these two types of adjectives interact with context. The section concludes with a discussion of the existing theoretical accounts of these differences. In section 2.3, I argue that given the fact that the theories presented in §2.2 are able to derive the correct interpretations for all gradable adjectives, it is not possible to adjudicate among them by exclusively relying on truth-conditional judgements. I propose that turning to online processing, and in particular to experimental paradigms that have been fruitful in studying matters pertaining to the interaction between context and interpretation, can

provide the relevant empirical evidence to settle this question. One such experimental paradigm is the Visual World Paradigm (Cooper 1974; Tanenhaus *et al.* 1995; Allopenna *et al.* 1998), where participant's eye-movements are tracked while they look at displays of objects and listen to linguistic instructions. Eye-movements have been successfully used as a dependent measure to investigate both language comprehension (see references above) and production (Meyer *et al.* 1998; Griffin & Bock 2000). More important for us, this paradigm offers a window into these processes by revealing with millisecond granularity when non-linguistic information from the visual context is integrated during language processing. Therefore, the Visual World paradigm presents itself as an optimal method to investigate the question of whether relative and absolute adjectives interact with context in comparable ways or not. In section 2.3.1, I review previous work that has made use of the Visual World paradigm to study the processing of prenominal adjectives during online reference resolution, as these constitute the base for Experiment 1, presented in chapter 3. Finally, in Section 2.4, I conclude.

2.2 Gradable adjectives and varieties of context-sensitivity: vagueness vs. imprecision

2.2.1 Gradability in the adjectival domain

Gradable adjectives such as *tall*, *full* or *dirty* (Creswell 1976; Cruse 1980; von Setechow 1984; Bierwisch 1989; Yoon 1996; Heim 2000; Kennedy & McNally 2005, Kennedy 2007, among many others) are adjectives whose meaning encodes an ordering along at least one dimension. In the case of *tall* such dimension is height, whereas in the case of *full* such dimension is fullness. Non-gradable adjectives, on the other hand, do not have scalar meanings; instead, they encode binary properties. Consider for instance the non-gradable adjective *atomic*. For any given individual, it is either the case that it is atomic or not. More importantly, all atomic individuals have the same *amount* of the property. Gradable adjectives, on the other hand, present a more complex picture, as

the properties denoted by this class of adjectives can be instantiated to different degrees, e.g. there exist infinite heights. Furthermore, it is even possible for an individual to have some amount of the relevant gradable adjectival property without qualifying as having the property itself. For instance a 5 feet tall adult man has height but intuitively does not qualify as tall.

Linguistic evidence that gradable adjectives have scalar meanings, unlike non-gradable adjectives, comes from distributional facts pertaining to degree modifiers such as *very*. As shown by the contrast in (10)-(11), degree modifiers like *very* can only be combined felicitously with gradable adjectives (Bolinger 1972; Rotstein & Winter 2004; Kennedy & McNally 2005). This distributional constraint arises from the fact that degree modifiers target a scalar ordering that is unavailable for non-gradable adjectives, hence the ill-formedness of (11).

(10) Gradable Adjectives

a. John is *very* tall.

b. This compound is *very* radioactive.

(11) Non-Gradable Adjectives

a. ???The clock is *very* atomic.

b. ???Dinosaurs are *very* extinct.

Comparative constructions provide a second piece of evidence for the gradable vs. non-gradable distinction. A comparative establishes an asymmetric relationship with respect to the degree that each of the two individuals being compared possesses the adjectival property. Under the assumption that non-gradable adjectives do not have scalar meanings, it is expected that only gradable adjectives should be able to participate in comparatives. As seen in (12) and (13), this contrast is borne out.

(12) Gradable Adjectives

a. John is *taller* than Vince.

b. Nuclear waste is *more radioactive* than coal ash.

(13) Non-Gradable Adjectives

a. ???The Hiroshima bomb was more atomic than the Nagasaki bomb.

b. ???Dinosaurs are *more extinct* than dodos.

An important concept related to the gradable vs. non-gradable distinction is the notion of *threshold* or standard of comparison. It has been proposed that the meaning of a gradable adjective involves a threshold whose value is often implicit and can even vague (Bartsch & Vennemann 1973; Creswell 1976; Klein 1980; von Setechow 1984; Bierwisch 1989; Ludlow 1989; Kamp 1985; Lewis 1979; Kennedy 1999; Kennedy & McNally 2005; Kennedy 2007; Graff 2000; Lassiter & Goodman 2013, 2015, among many other). Under this view, sentence (10b) asserts that John's height exceeds a threshold of height above which those individuals belonging to the same category as John, e.g. adult males, are considered to be tall. The meaning of non-gradable adjectives does not involve thresholds because non-gradable adjectives do not encode an ordering along an adjectival scale, and so there are no relevant distinctions to be made among different degrees of the relevant adjectival property.

More importantly, the notion of threshold explains why gradable adjectives do not form a completely homogeneous class. Kennedy (2007) observes that gradable adjectives can give rise to three different types of entailment patterns when used in the comparative construction. In comparatives of the form *X is more A than Y*, some gradable adjectives like *wet* entail that X is A (14); some adjectives like *dry* entail that B is not A (15); and finally adjectives like *big* entail neither that X is (not) A nor that Y is (not) A (16).

(14) Minimum Standard Absolute Adjectives

a. The red towel is wetter than the blue towel.⇒

b. The red towel is wet.

(15) Maximum Standard Absolute Adjectives

a. The red towel is drier than the blue towel. ⇒

b. The blue towel is not dry.

(16) Relative Adjectives

a. The red towel is bigger than the blue towel. ⇒

b. The red towel is (not) big.

c. The blue towel is (not) big.

The three inference patterns exemplified in (14)-(16) have motivated a classification of gradable adjectives into Maximum and Minimum Standard Absolute Adjectives (14)-(15), and Relative Adjectives (16). The labels respond to the intuition that Minimum Standard Absolute Adjectives (henceforth MinAAs) require that an individual possess only a *minimal* degree of the property for the adjective to be truthfully predicated of such individual. Thus, the threshold of MinAAs coincides with to the minimal degree of the adjectival scale. Maximum Standard Absolute Adjectives (henceforth MaxAAs) require the opposite; for an object to possess the property denoted by a MaxAA, it must be the case that the object possesses the *maximal* degree of the adjectival property.[1] Therefore, in the case of MaxAAs, the value of the threshold corresponds to the maximal degree of the adjectival scale. In the case of Relative Adjectives (henceforth RelAAs), the threshold can correspond to any degree of the adjectival scale. As discussed earlier in this section, the degree of height of a tall jockey differs greatly from the height of tall NBA player. This is the reason why it is possible to utter a sentence like (17) without incurring in a contradiction; At an intuitive level, sentence (17) is acceptable because the heights of John and Pete are being evaluated with respect to two different classes of individuals, namely jockeys and NBA players respectively.

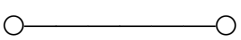
(17) Despite having the same height, John, the Jockey, is tall, while Pete, the NBA player, is not.

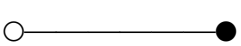
The class used to evaluate whether a certain individual possesses a gradable property is often referred to as the *Comparison Class* (Kamp 1985; Klein 1980; Kennedy & McNally 2005;

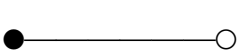
1. Although see the discussion below for a qualification of this claim.

Kennedy 2007; van Rooij 2011; Sassoon & Toledo 2011; Burnett 2014). While it is clear that relative thresholds are always computed with respect to a contextually salient Comparison Class, it has been argued that the thresholds of MinAAs and MaxAAs are not relativized to a Comparison Class, but rather have *absolute* values corresponding to either the minimum or maximum degree of the scale. Hence, MinAAs and MaxAAs are often conflated and referred to as Absolute Adjectives (AAs).

The different properties of relative and absolute thresholds have been traced back to the structural characteristics of the scales used by each of these classes of adjectives. Kennedy & McNally (2005) propose that the scales of RelAs are open ended in both extremes of the scale (18a), while AAs have at least one bound endpoint (18b-c). Lower-bound scales give rise to MinAAs meanings (18b), whereas upper-bound scales give rise to MaxAAs meanings (18c).

(18) a.  RelAs

b.  MaxAAs

c.  MinAAs

The distribution of modifiers like *slightly* and *completely* has been claimed to diagnose minimum and maximum scalar points respectively (Kennedy & McNally 2005; Rotstein & Winter 2004; Sassoon & Toledo 2011, among others).[2] As seen in (19), the modifier *slightly* is only compatible with adjectives whose scale is closed in the lower end, thus diagnosing MinAAs, while modifiers like *perfectly* can only be combined with adjectives that have an upper-bound scale, thus diagnosing MaxAAs.

(19) a. ?slightly/ ?perfectly tall.

b. slightly/ ?perfectly dirty.

c. ?slightly/ perfectly clean.

2. Strictly speaking, these modifiers test for minimum and maximum scalar endpoints, respectively, which are independent of — though generally correlated with — maximum and minimum standards.

Going back to the entailment patterns in (14)-(16), the fact that RelAs do not trigger any entailments regarding whether the adjectival property holds of either the target or the standard of comparison follows straightforwardly from the fact that they have relative thresholds; the mere fact that one object exceeds another with respect to some relative property tells us nothing about how the objects stand in relation to a contextually salient comparison class. On the other hand, if the value of absolute thresholds defaults to the the scalar endpoint value, it follows that MaxAAs should trigger the entailment that the standard of comparison does not possess the adjectival property. By the same token, assuming that minimum standard thresholds coincide with the lower-bound scalar point, it is correctly predicted that this class of adjectives used in the comparative should give rise to the entailment that the target of comparison holds the property denoted by the MinAA.

The question of how to model non-GAs seems straightforward. From a set theoretic perspective, non-gradable adjectives can be easily modeled as sets of individuals (e.g. the adjective *atomic* denotes the set of all atomic individuals). However, for all the reasons discussed above, gradable adjectives usually receive a more complex denotation (to be discussed below), since the meaning of gradable adjectives is highly sensitive to the class of individuals that the adjective describes (e.g., compared to a tall NBA player, a tall jockey is actually very short). Therefore, it is clear that the denotation of an adjective like *tall* should not consist of the set of *all* tall individuals.

Within the degree semantics framework (Creswell 1976, von Setechow 1984, Kennedy 1999; Heim 2000; Kennedy & McNally 2005; Kennedy 2007; Syrett *et al.* 2009; Solt & Gotzner 2012, among many others), built on the assumption that degrees (type d) are a basic category of the model-theoretic ontology, GAs have been formalized as relations between a degree and an individual (20).[3]

3. It should be noted that there also exist degreeless alternatives to the formalization of GAs. In particular, Delineation Semantics accounts (Klein 1980; Doetjes *et al.* 2009; Burnett 2014, among others) model adjectival predicates as sets of individuals that are always evaluated with respect to a particular, contextually provided, comparison class. In this dissertation, I will assume degrees as a basic semantic type, although this decision does not have any bearing on the set of research question that the dissertation seeks to address.

(20) $[\![\text{ A }]\!] = \lambda\theta_A \lambda x[\mu_{A(x)} \geq \theta_A]$

More specifically, the denotation in (20) defines the meaning of a GA as an asymmetric relation between the value of the free variable θ, corresponding to the adjective threshold, and the degree to which the individual instantiates the adjective property (where $\mu_A(x)$ is the measure of x in the adjective scale A, and A denotes an ordered set of degrees). As it is, the denotation in (20) does not allow for direct composition with an individual. In degree semantics this compositional problem is fixed by posing a degree morpheme *POS* (Creswell 1976; von Setechow 1984; Kennedy 1999; Kennedy & McNally 2005; Kennedy 2007; Grano 2012) that provides a free variable (S_A in (21a)), and whose value is resolved contextually. This silent degree morpheme combines directly with the adjective, binding its threshold argument, as seen in (21b):[4]

(21) a. $[\![POS]\!] = \lambda A \lambda x[A(s_A)(x)]$

b. $[\![POS\ A]\!] = \lambda x[\mu_A(x) \geq s_A]$

In (21), the value of the threshold free variable s_A is always resolved via context. While it is clear that RelAs have context-sensitive thresholds, it is a question of debate whether the same mechanism extends to the thresholds of AAs, or whether in the latter case the threshold value is grammatically encoded in the lexical semantics of the absolute predicate as in (22a-b).

(22) a. $[\![POS_{max}\ A]\!] = \lambda x[\mu_A(x) \geq \mathbf{max}_A]$

b. $[\![POS_{min}\ A]\!] = \lambda x[\mu_A(x) \geq \mathbf{min}_A]$

In the remainder of this dissertation, I will make a distinction between theories that pose *shifting semantic thresholds* for both RelAs and AAs, and theories that assume variable thresholds of RelAs, but pose *fixed thresholds* for AAs. In section 2.2.3, I discuss these two approaches in greater detail.

4. Alternatively, *POS* can also be treated as a type shifter that changes the order of the arguments in (20).

2.2.2 Context-sensitive interpretations of Gradable Adjectives

Gradable relative (23a) and absolute (23b) adjectives in their positive form have been claimed to have context-sensitive uses (Unger 1975; Kennedy & McNally 2005; Kennedy 2007; McNally 2011, among many others).

(23) a. John is *tall*.

b. The movie theater is *empty*.

RelAAs are context-sensitive because their criteria of application change across contexts: as discussed in the preceding section, different contexts pick out different Comparison Class in order to determine what subset of individuals in the Comparison Class qualifies as having the adjectival property. This can be observed in example (23a), where the threshold that distinguishes those individuals that count as tall from those that do not varies as a function of the features of the context. For instance, the cut-off point determining what individuals fall in the extension of the predicate *tall* will be set to a higher value in a context in which the height of basketball players is discussed. In contrast, in a context where the question under discussion is the height of male jockey (who tend to be much shorter than male basketball players), the cut-off point will be significantly lower. Thus, if John is 5'10ft tall, (23a) is true in case John is a jockey, but false in case he is a basketball player.

The criteria of applicability of AAs can also vary from context to context. Clearly, sentence (23b) can be truthfully uttered in a context in which there are no people in the movie theater. Intuitively, it also seems possible to felicitously utter (23b) in a context where there are only two people sitting in the back row of the theater during the premier of a popular movie. In such a context, the fact that the movie theater is not *completely empty* can be ignored for the purpose of uttering (23b). Other contexts, however, impose further restrictions on the amount of available information that can be ignored without rendering an utterance of (23b) infelicitous. In the same situation where there are two people sitting in the back row, example (23b) stops being acceptable if uttered to

describe the movie-theater during a fire-emergency that requires the building to be evacuated. In this context, it seems clear that the two people sitting in the back of the room can no longer be ignored. Unlike the previous context, in such a context *empty* must mean *completely empty*.

The latter discussion shows that the context-sensibility displayed by AAs is more restricted than that observed with RelAs. While the cut-off points, or thresholds, of RelAs display a great deal of variability (e.g. it is possible to refer to a toddler as tall, even though he is possibly much shorter than a short adult), thresholds of AAs are more rigid, tending to be interpreted as endpoint oriented (e.g. in order to count as empty, the movie-theater must be—closed to—maximally empty). Put it differently, the contextual variability displayed by AAs seems to be reduced to how much distance from the endpoint-oriented interpretation of the threshold is tolerated in a given context.

Behavioral evidence confirms the intuition that RelAs and AAs show different degrees of context-sensitivity. In an acquisition study comparing children and adult controls, Syrett *et al.* (2009) show that the latter group consider the positive form of a RelA to be felicitous when used to distinguish between two objects that have been previously judged to not be adequately described by that same adjective. For instance, when presented with two rods of different length, neither of which was judged to be long separately, participants accepted the request in (24a) and provided the experimenter with the longer rod in the set up, showing that they were willing to accommodate the existence presupposition of the definite article '*the*' (namely, that there existed a unique long rod in the context).

(24) a. Please, give me the *long* one. [Relative adjective trial]

b. Please, give me the *full* one. [Absolute adjective trial]

While adults agreed to perform the task in trials containing a RelA, they consistently rejected to do so in trials containing AAs. For instance, when adult participants were presented with two partially full jars filled to different degrees (see Figure 2.1), they rejected the request in (24b).

The contrast found by Syrett *et al.* (2009) between RelAs and AAs is consistent with the claim that RelA's thresholds are more malleable than those of AAs. The value of a RelA threshold can be

set to any value in order to satisfy the demands of the context, or, put it differently, its value is always contingent on the properties of the context. Thresholds of AAs, on the other hand, are to some degree context-independent; the use of an AA ceases to be acceptable once the context requires to set the threshold too far away from the scalar endpoint, thus explaining why adult participants rejected to perform the task in AA trials.

Figure 2.1: Experimental set up used by Syrett *et al.* (2009).

A second distinction between RelAs and AAs concerns the phenomenon of *vagueness*. Even though the literature contains many characterizations of what it entails to be a vague predicate (Kennedy 2007; van Rooij 2011; Burnett 2014, among many others), there exists general consensus that a core property of vague predicates is their ability to give rise to *borderline cases*, i.e. individuals for which it is not possible to determine whether they are in the extension of the adjective or not (Bonini *et al.* 1999; Alxatib & Pelletier 2009). For instance, John is a borderline case with respect to the adjective *tall* if it is not clear whether he could be truthfully described by the adjective or not. A good test for determining whether a predicate A has borderline cases is to plug it in sentences whose classical logic representation would constitute a contradiction, e.g. $\neg A(x) \wedge A(x)$ or the logically equivalent $\neg(A(x) \vee \neg A(x))$. This is exemplified in (25a) for the relative adjective *tall*. The acceptability of (25a) shows that the predicate *tall* can give rise to borderline cases. As expected, a non-gradable predicate, like *extinct*, which is not vague and hence does not give rise to borderline cases, fails this test as shown by the contradictory status of (25b):

(25) a. John is neither tall, nor not tall.

b. #The Javan rhinoceros is neither extinct, nor not extinct.

Whether or not a predicate has borderline cases determines the way in which it interacts with context. There exists general agreement that RelAs have borderline cases irrespectively of the properties of the particular context of evaluation. However, Kennedy (2007) notes that for AAs, it is always possible to find a context that will make fine-grained enough distinctions such that any potential borderline case will be settled. Sentence (23b) repeated below in (26) offers such an example. In the context of a famous movie premier attended by only twenty people, the movie-theater could be construed as a borderline case with respect to the predicate *empty*. After all, in such scenario there is a significant degree of epistemic uncertainty as to how many people is necessary for the movie-theater to stop counting as empty. However, it is always possible to construct scenarios that would make clear cut distinctions with respect to the applicability of the predicate *empty*, thus eliminating any potential borderline cases. For instance, as discussed above, in the context of an emergency evacuation during a fire, it is clear that there is only one situation in which (26) can be used felicitously, namely when there is no one in the building.

(26) The movie theater is *empty*.

Based on this type of contrasts, Kennedy argues that absolute adjectives allow for *natural precisifications* (Pinkal 1995).[5] Relative adjectives, on the other hand, do not. Consider the contrast in (27), discussed by Kennedy (2007), involving the relative adjective *long* and the absolute adjective *straight*.

(27) a. ??We need a long rod for the antenna, but since *long* means 'greater than 10 meters' and this one is 1 millimeter short of 10 meters, unfortunately it won't work.

5. This distinction is also at the base of another difference between RelAs and AAs, namely that the Sorites Paradox is less compelling with AAs than RelAs (Kennedy 2007). See Burns (2012) for a recent overview of the solutions provided to this paradox.) This is due to the fact that the second inductive premise seems less acceptable in the case of AAs than RelAs, precisely because only the former allow for natural precisifications.

b. The rod for the antenna needs to be *straight*, but this one has a 1 mm bent in the middle, so unfortunately it won't work.

Example (27a) shows that a small difference of 1 millimeter does not seem to be sufficient to differentiate those rods that are long from those that are not, even in cases in which the adjective threshold is made explicit.[6] This suggests that the threshold of the positive form of a RelA is fairly insensitive to small degrees of change (e.g. going from a 10 meters long rod to one that is 9.999 meters long is not sufficient to stop counting the rod as long). Example (27b), on the other hand, shows that this is not the case with AAs, for which it is always possible to find a more precise context in which a small degree of change (e.g. 1 mm) will be sufficient to differentiate those individuals that fall in the extension of the predicate from those that do not.

Another domain where parallel differences between RelAs and AAs emerge is in the availability of *crisp judgements* in implicit comparison constructions such as (28a). The implicit comparison construction requires that an ordering is established between two objects x and y with respect to a gradable property A by using the positive form of A such that A is true of x and false of y. Crisp judgement effects arise when a particular expression cannot be used to describe differences of a very small degree. Our previous discussion predicts that RelAs should give rise to crisp judgement effects, whereas AAs should not, as they can be precisified. Sentence (28a) is intended as a description of the Panel A in Figure 2.2, whereas sentence (28b) is intended as a description of Panel B. Panel A represents two lines, one of which is completely straight (line B), whereas the second one is slightly curved (line B). Panel B represents two cylinders, both of which are short, although cylinder B is slightly shorter than cylinder A. In each case, the difference in the degree to which the two objects have the relevant adjectival property, i.e. height vs. straightness, is small. Importantly,

6. An open question is whether the value of the threshold of a RelAs can be made precise to begin with, as in (27a). The oddity of (27a) could also be explained if RelAs thresholds are vague by definition, i.e. always involve uncertainty. If relative standards, unlike absolute standards, are always uncertain, the existence of borderline cases with RelAs would follow. The relevant point for us is that the unavailability of natural precisifications with RelAs (whatever their source might be) has consequences for the ways this type of predicates interact with context.

the contrast in acceptability shown in (28a-b) shows that in contexts requiring crisp judgements, implicit comparison is unacceptable with RelAs, whereas AAs allow for such fine-grained distinctions.

Figure 2.2: **Left:** Panel A; **Right:** Panel B

(28) a. Compared to line A, line B is straight.

b. #Compared to cylinder A, Cylinder B is short.

After having laid out the empirical differences between RelAs and AAs with respect to their context-sensitivity, the following section describes the existing theories tailored to account for the properties of GAs discussed in this section.

2.2.3 Accounts of the Relative vs. Absolute distinction

Given that RelAs show unrestricted contextual variability, existing accounts of the relative/absolute distinction assume that the threshold of RelAs is always set by accessing contextual information. However, theories differ with respect to the status they assign to absolute thresholds. Broadly speaking, proposals can be split based on whether they take the semantic threshold of relative *and* absolute adjectives to be variable or whether they assume that absolute thresholds are fixed by semantic content.

Shifting absolute thresholds

Among the theories that argue for variable absolute thresholds is Lassiter & Goodman (2013), who propose an account that makes use of the Bayesian Rational Speech Act framework (Frank & Goodman 2012; Bergen *et al.* 2012; Frank & Goodman 2014; Zeevat 2014; Goodman & Stuhlmüller

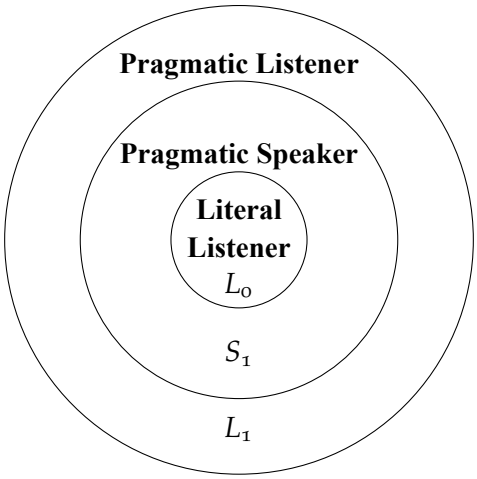


Figure 2.3: Model of the pragmatic listener (L_1).

2013; Kao *et al.* 2014; Lassiter & Goodman 2015; Qing & Franke 2014; Potts *et al.* 2016; Goodman & Frank 2016, among many others). Before describing Lassiter & Goodman (2013)'s theory, I first start by providing a short introduction to the RSA framework.[7]

RSA models treat language interpretation as a rational probabilistic inference process in which speakers and hearers recursively reason about each other. Modeling language understanding as Bayesian inference, RSA models offer a straightforward way to derive pragmatically enriched interpretations that go beyond the strict truth-conditional content of an utterance. RSA models formalize this pragmatic listener as a Bayesian agent who, in order to arrive at the utterance interpretation intended by the speaker, reasons about an idealized pragmatic speaker who, in its turn, reasons about a naïve literal listener. This recursive reasoning chain is graphically depicted in Figure 2.3.

In the pragmatic listener model exemplified in Figure 2.3, the **literal listener** (L_0) as formalized in (29) infers the state of the world *s* by uniquely conditioning on the literal interpretation of utterance u. Therefore, this layer of the model corresponds to the point in which the semantic truth-conditions of *u* are factored in during the process of pragmatic interpretation.

7. The following overview is mostly based on Lassiter & Goodman (2013, 2015) and Scontras & Tessler (2017).

(29) $P_{L_0}(s \mid u) \propto P_{L_0}(s \mid [\![u]\!] = 1)$

The next layer corresponds to the **pragmatic speaker** (S_1). This second layer formalizes the process by means of which the pragmatic speaker chooses her utterance among a set of possible alternatives. Utterance choice is constrained by two factors: 1) what utterance has the best chances of conveying the state of the world *s* to the literal listener; and 2) the pragmatic speaker's private preference of choosing an utterance that minimizes production effort.[8] Based on this two criteria, the speaker ranks possible utterances by means of the utility function defined in (30), which precisely captures this tension. The utility of an utterance *u* and a state *s* is proportional to how informative *u* is to the literal listener L_0, i.e. how effective *u* is at communicating the state of the world *u* to L_0, minus the utterance cost incurred by *u*. In these models, informativity is often cast as negative surprisal (positive log probability, Shannon (1948)) of the state of the world *s* once the literal listener has conditioned on the truth of the utterance. The lower the surprisal of *s* given *u*, the more predictable and therefore more informative *u* becomes for the literal listener. Utterance cost, on the other hand, is usually quantified as utterance length: the longer the utterance, the higher the cost. Based on this, an utterance *u* will more efficiently communicate *s* to the literal listener the lower its surprisal and utterance cost.

(30) $U_{S_1}(u;s) = log(P_{L_0}(s \mid u)) - C(u)$

Given that the pragmatic speaker chooses her utterance by maximazing its utility, the probability that S_1 will choose utterance *u* given the state *s* is proportional (30). This is captured in the model of the pragmatic speaker in (31). Importantly, to make speaker model more cognitively plausible and reflect the fact that speakers do not *always* choose their utterances rationally, utterance choice is governed by some degree of stochasticity determined by the term $\alpha > 0$ in (31). The bigger the value of α the more the pragmatic speaker approximates the deterministic choice based on (30).

8. At this point, the question of what precise internal factors shape the speaker's preferences is an open question. The models presented in this section make minimal assumptions about such preferences, so this aspect of the models, and whether the preferences used are sufficient or well motivated will not be discussed any further.

(31) $P_{S_1}(u \mid s) \propto exp(\alpha U_{S_1}(u;s))$

The last layer of the model corresponds to the **pragmatic listener** (L_1). As shown in (32), the pragmatic listener infers the state of the world *s* given that the speaker has uttered u. This bayesian inference is performed by taking into consideration the likelihood that S_1 would utter u given that she intends to convey *s*, as well as the prior probability that the pragmatic listener assigns to *s* being the case.

(32) $P_{L_1}(s \mid u) \propto P_{S_1}(u \mid s) \cdot P_{L_1}(s)$

We are now in a position to introduce Lassiter & Goodman's RSA model for the interpretation of gradable adjectives. Lassiter & Goodman assume a unified denotation for all classes of gradable adjectives (33).

(33) $[\![A]\!] = \lambda\theta_A \lambda x[\mu_A(x) > \theta_A]$

Like other degree accounts, the denotation in (33) contains a free variable over thresholds. Crucially, the denotation in (33) assumes that *both* RelAs and AAs have variable thresholds whose values are fixed in context. The generic RSA model described above does not include a mechanism for performing bayesian inference over free variables. More specifically, the literal listener as defined in (29) cannot condition over the truth of an utterance whose semantic truth-conditions contain a free variable. This problem is overcome by having the pragmatic listener infer not only the state of the world *s* given u, but also the value of the adjectival threshold as parametrized by the assigned by the function V (34-37). Technically, this is implemented by passing the free variable all the way up to the level of the pragmatic listener, who considers all possible assignment values V to the threshold free variable (37).

(34) $P_{L_0}(s \mid u, V) \propto P_{L_0}(s \mid [\![u]\!]^V = 1)$

(35) $P_{S_1}(u \mid s, V) \propto exp(\alpha U_{S_1}(u;s,V))$

(36) $U_{S_1}(u;s,V) = log(P_{L_0}(s \mid u, V)) - C(u)$

(37) $\mathrm{P}_{L_1}(s, V \mid u) \propto P_{S_1}(u \mid s, V) \cdot P_{L_1}(s) \cdot P_{L_1(V)}$[9]

The formalization of the pragmatic speaker in (37) deserves a more detailed commentary. First, there exists an underlying assumption that the prior over possible values of *s* corresponds to a prior on the relevant comparison class in the context of utterance. For instance, for the utterance *John is tall* and given that John is an adult man, the term $P_{L_1}(s)$ in (37) would consist of a probability distribution over heights of adult men as determined by the listener's world knowledge.[10] Second, it is assumed that hearers do not make use of prior world knowledge about likely threshold values in the interpretation of gradable adjectives. This assumption is captured by letting $P_{L_1(V)}$ consist of a flat probability distribution where all possible threshold values are equally likely. Rather, the joint posterior distribution of *s* and *V* results from a trade off between the speaker's preference for informativity and the listener's believes about the probability of *s* being the case. Very high threshold values, e.g. 7 feet tall, are very informative but have a very low probability, i.e. few adult males are taller than 7 feet. On the other hand, a low threshold, e.g. 5 feet tall, would have a very high probability but would not be very informative, since most adult men are taller than 5 feet. Therefore, the pragmatic speaker assigns higher posterior probabilities to threshold values and state pairs that are both informative given the utterance and probable given previous world knowledge.

As described in (33)-(37), the model designed by Lassiter & Goodman makes no reference to specific subclasses of gradable adjectives. This is in fact an important feature of this proposal: the relative vs. absolute distinction is not grammatically grounded in the sense that it is not derived from core semantic properties of the predicates. In this system, the relative vs. absolute distinction arises from prior world knowledge about the statistical properties of the comparison class used for the interpretation of a gradable adjective. Relative-like interpretations arise when the prior distribution is approximately normal (2.4), whereas absolute-like interpretations, result from skewed priors (see

9. $P_{L_1(V)}$ is a constant and can therefore be dropped.

10. For an extended model in which the pragmatic listener also infers the correct comparison class, see Tessler *et al.* (2017).

Figure 2.5).

As can be seen in Figure 2.4, evaluating the adjective *tall* in a sentence like *John is tall* against a quasi normal prior over heights results in a height posterior that is also roughly normally distributed, although more peaked.[11] Crucially the height posterior is pushed to the right to reflect the fact that we are now considering heights above average. Figure 2.5 contains the simulation results for the antonym pairs *dangerous*/*safe*.[12] The Beta distribution used to model the danger prior in Figure 2.5 is skewed toward the 0 point, with almost all the probability mass accumulating in the low values of the degree of danger scale. This is supposed to capture the intuition that less dangerous events are considered to be more likely than very dangerous ones. As can be seen in the left panel of Figure 2.5, the danger posterior resulting from the pragmatic interpretation of the adjective *dangerous* is such that the probability of something counting as dangerous rapidly increases as it gets away from zero, although very high degrees of danger will still be considered unlikely. On the other hand, in the case of *safe* (right panel of Figure 2.5), the danger posterior distribution drastically decreases as it gets away from zero, reflecting that only events that are maximally, or close to maximally, safe can be felicitously described by the adjective *safe*.

This account of the relative vs. absolute distinction can explain the observation that context-sensitivity of AAs is more restricted compared to RelAs, as the posterior probability distribution for possible threshold values presents a much smaller standard deviation when the adjective prior is skewed, compared to when it is roughly normal. Since likely values for the threshold are fairly

11. Lassiter & Goodman make the simplifying assumption that the alternative utterances considered during the interpretation of a sentence like *John is tall* are *John is short* or saying nothing. The question of how to generate the relevant alternative utterances, and how these alternative utterances constrain the possible range of interpretations of *u* has not been properly investigated, although see Peloquin & Frank (2016) for a preliminary investigation of this question using the case study of scalar implicatures.

12. Gradable adjective antonym pairs are assumed to be represented in the same lexical scale but with opposite polarities. More generally, this entails that the only difference in the lexical representation of an antonym gradable adjective pair is in the direction of the greater-than relation in the lexical entry specified in (33).

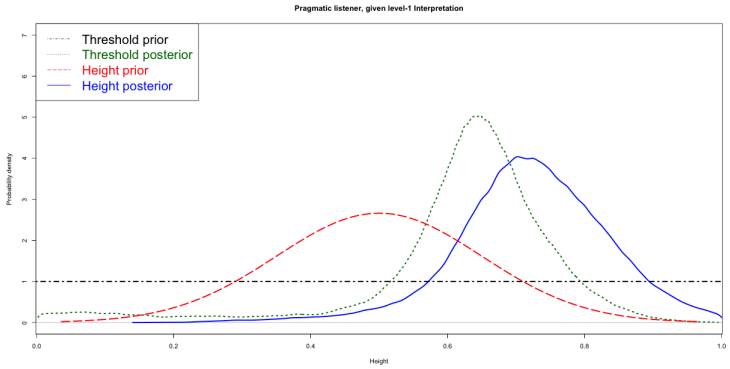


Figure 2.4: Simulation results for the RelA *tall* as presented in Lassiter & Goodman (2013) using an approximately normal prior of adult male heights.

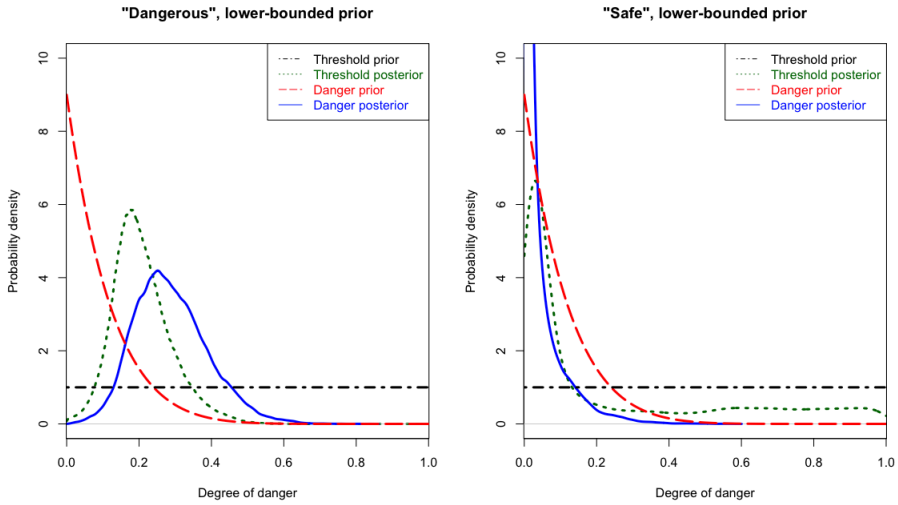


Figure 2.5: Simulation results for the MinAA and MaxAA antonym pair *dangerous*/*safe* as presented in Lassiter & Goodman (2013) using a skewed *danger* prior.

restricted, it is correctly predicted that absolute adjectives should give rise to little uncertainty about the threshold value, thus leaving less room for context-driven variability.

Fixed absolute thresholds

The second type of approaches assume that absolute thresholds are not subject to contextual variability: truth-conditionally, AAs always have endpoint-oriented meanings. Context-sensitive uses of AAs are claimed to emerge from pragmatic reasoning about *imprecision* (Kennedy 2007; Syrett *et al.* 2009; van Rooij 2011; Burnett 2014; Qing & Franke 2014; Leffel *et al.* 2016). Here, I focus on one possible implementation of this view, that of Kennedy (2007), introduced in §2.2.1. Building on previous work on the topology of adjectival scales (Kennedy & McNally 2005), Kennedy (2007) traces the relative/absolute distinction back to the structural properties of the scale associated with the adjective. Kennedy proposes that the unrestricted context-sensitivity displayed by RelAs derives from the fact that their scales are open. Since the scales associated with RelAs are unbounded in at least one of its end-points, the value of the threshold cannot be provided by the scale itself and in most cases must be computed with respect to a contextually salient Comparison Class. AAs, on the other hand, are associated with closed scales, whose endpoints provide a conventional value for the adjectival threshold. Therefore, in this framework, the context-sensitivity displayed by RelAs and AAs is fundamentally different: RelAs are semantically context-sensitive, whereas AAs are pragmatically context-sensitive.

Kennedy (2007) does not provide an explicit theory of imprecision. However, there exist compatible approaches to this phenomenon that can derive non-endpoint-oriented standard interpretations for AAs without having to assume shiftable thresholds in the semantics. One such approach is Lasersohn's (1999) *Pragmatic Halo* theory of imprecision. Lasersohn discusses examples like (38), and points out that it would be very natural for a speaker to utter (38), even if a small percentage of the townspeople were awake.

(38) The townspeople are asleep.

In such scenario, Lasersohn argues, sentence (38) is truth-conditionally false, but close enough to the truth to still be pragmatically acceptable and informative. Lasersohn refers to these trivial deviations from the truth as the *pragmatic slack* of a linguistic expression. The innovation of this proposal is that any linguistic expression $[\![\alpha]\!]^{M,C}$ interpreted with respect to a model M and a context C is coupled with its *pragmatic halo*, which consists of the partially ordered pair $\langle H_C(\alpha), \leq_{\alpha,C} \rangle$. $H_C(\alpha)$ returns the elements that conform the pragmatic halo of α, including the denotation of α itself. The elements in the halo are always objects of the same semantic type as α that only differ from $[\![\alpha]\!]^{M,C}$ in pragmatically ignorable ways in C. The second element of the ordered pair, $\leq_{\alpha,C}$, is an ordering of $H_C(\alpha)$ based on its similarity to $[\![\alpha]\!]^{M,C}$. Halos are pragmatic because they are represented separately from the semantic truth-conditional content of the linguistic expression. However, an important feature of this model is that imprecise meanings are compositionally derived. The main motivation for making pragmatic halos accessible to the compositional semantics is that they interact with operators such as *exactly*, whose semantic contribution is to contract the halo of the expression they combine with (39). Lasersohn refers to these operators as *slack regulators*.

(39) Mary arrived at *exactly* three o'clock.

Finally, Lasersohn also adds a felicity condition on assertion that ensures that a sentence may be felicitously uttered if some element in its halo is true, even if the strict meaning of the sentence is not. Despite claiming that imprecision is a purely pragmatic phenomenon, the biggest virtue of Lasersohn's proposal is that it models the more conventional aspects of imprecision, especially with respect to the compositional effects of slack regulation. However, little is said regarding the pragmatic principles that regulate loose talk (see Lauer 2012 for a similar remark). Even though Lasersohn's goal was never to address the pragmatics of imprecision, ideally a full account of imprecision would have something to say about when and why speakers choose to be imprecise, as well as what kind of information hearers resort to in deciding to interpret a particular utterance imprecisely.

A second approach to imprecision is represented by the scale of granularity theory of imprecision, developed initially by Krifka (2002) (but see also Krifka 2007 and Krifka 2009) to model imprecision in the numeral domain.[13] Granularity-based accounts assume that any scale can be conceptualized with different levels of granularity that are related to each other by an homomorphic mapping. The level of granularity is determined, among other things, by the amount of distinctions each context makes among the degrees that form the relevant scale. In the adjectival domain the relevant scale would correspond to the scale provided by the adjective. For instance, in the imprecise interpretation of sentence (40a) a coarse adjectival scale would be at play such that two people in the room would be mapped to the maximal degree of emptiness. A more precise context imposing a finer-grained granularity level would distinguish between zero and two people in the room, such that only the former would be mapped to the maximal degree of emptiness.

(40) a. The theater is *empty*.

b. Fact: There were two people sitting in the back row of the theater.

Adjudicating between the two theories of imprecision introduced in this section is beyond the scope of this dissertation. The important point is that any of the two approaches to imprecision discussed above could be easily coupled with any of the theories that assume rigid semantic thresholds for AAs, such as Kennedy's proposal, in order to explain why context-sensitive interpretations of AAs are available in the first place.

Context-sensitive interpretations of MinAAs

It should be pointed out that the context-sensitivity displayed by MinAAs cannot actually be modeled as imprecision. Despite surface similarity, the logic underlying context-sensitive interpretations of MinAAs and MaxAAs is diametrically opposed.[14] Context-sensitive interpretations of

13. Granularity accounts of imprecision will be presented in more depth in Chapter 4, where I discuss context-sensitive interpretations of number words.

14. I thank Ming Xiang for pointing out this difference between MinAAs and MaxAAs to me.

MaxAAs involve weakening of the semantic truth-conditions (i.e. the non endpoint-oriented interpretation of the adjective threshold is asymmetrically entailed by the semantic denotation), while context-sensitive interpretations of MinAAs involve strengthening of the semantic truth-conditions (i.e. the semantic truth-conditions are asymmetrically entailed by the non endpoint-oriented interpretation of the adjective threshold).[15]

A second difference between the context-sensitivity of MaxAAs and MinAAs lies in the ways in which such uses emerge in conversation. While context-sensitive uses of a MaxAA can be easily identified from the interpreter's perspective, at least in those cases that involve referential communication, it is virtually impossible for the comprehender to identify a context-sensitive use of a MinAA, i.e. to determine whether the speaker intended her utterance of a MinAA to be interpreted with a non-minimal threshold. In fact, the only instance in which speakers and hearers can coordinate about the value of non-minimal thresholds is in situations in which this value is explicitly at stake, as in cases where the predication of a MinAA is rejected by the hearer, despite there being a non-minimal degree of the adjectival property. This is exemplified in (41), where Ben challenges John's use of the MinAA *dirty* to describe their living room, even though Ben can also see that there is some dirt lying on the floor.

(41) John and Ben spent the morning cleaning their apartment, but they did not have time to finish one of the kitchen cabinets. At night, while watching TV, John notices a spot of dust in one corner of the room and while pointing at it says:

a. John: This living room is dirty.

b. Ben: No, it's not! The living room is clean, but the kitchen is still dirty!

15. The characterization of context-sensitive interpretations of MaxAAs and MinAAs as *weakening* and *strengthening* of the semantic truth-conditions respectively only holds under the assumption that MinAAs and MaxAAs denote fixed end-point oriented thresholds. For theories that assume shifting thresholds across the board in the form of a free variable in the semantics, characterizing context-sensitive interpretations in terms of weakening or strengthening of the semantic truth-conditions is not possible because in this view all AAs are semantically context-sensitive, i.e. the meaning of an AA is always resolved by accessing contextual information.

c. John: Well, I guess you are right, the living room is clean, not dirty.

What examples like (41) show is that minimal degrees of the property can be judged to not be sufficient in order to license the use of a MinAA. Importantly, John's use of the MinAA *dirty* in (41a) is *not* context-sensitive, since it is perfectly compatible with an endpoint-oriented interpretation of the threshold. In fact, the only instance in (41) where the use of the adjective *dirty* can be claimed to be context-sensitive—in the sense that the adjective can receive a non-endpoint oriented interpretation, are cases involving the negation of the MinAA *dirty*, which are otherwise tantamount to an imprecise interpretation of the antonym MaxAA *clean* as shown by (41c).

From this perspective, there is not such a thing as a context-sensitive interpretation of a MinAA, since any felicitous utterance of a MinAA will necessarily be compatible with an endpoint-oriented interpretation of the threshold. Therefore, when it comes to MinAAs, considerations about context can determine whether the use of a MinAA will be licensed in the first place, but such considerations in no case will constraint the range of interpretations that the MinAA can receive once the adjective has been uttered by the speaker, and the utterance has been accepted by the hearer.

Assuming that there are no true context-sensitive utterances of MinAAs, the inferential task that the comprehender is faced with upon hearing a MinAA is simpler compared to the inference task that he needs to perform in order to arrive at the interpretation of either a RelA or a MaxAA. Since the interpretation of a non-negated MinAA is not dependent on context, the interpreter does not need to infer what threshold value was intended by the speaker. In other words, the interpreter can simply use *any* threshold value that would be sufficient to make the speaker's utterance true.[16] The crucial point is that this is not an inference requiring reasoning *about* the speaker's criteria of applicability of the MinAA. Rather, the interpreter uniquely needs to consult his own criteria of applicability in order to determine whether he accepts or not the utterance of the MinAA in the context. If he does not, he should then proceed to challenge the speaker's use of the MinAA much

16. At this point, I abstract away from the question of whether interpreters choose more conservatively to default to minimal threshold values across the board, or whether they adopt non-minimal thresholds as well.

in the same way that Ben does in example (41).

The two types of theories of the relative vs. absolute distinction that have been presented in this chapter cannot equally accommodate for the facts pertaining to MinAAs described in this section. In particular, Kennedy's account of gradable adjectives, can easily account for the insensitivity of MinAAs to context during interpretation, since it assumes that absolute thresholds are semantically set either to a minimum or to a maximum degree, and it is these truth-conditions what the interpreter resorts to for the interpretation of the adjective. Even though Kennedy himself does not discuss this point, world knowledge would only come into play to reject uses of the adjective for which the truth-conditions are too weak.

The issues discussed in this section are harder to integrate in Lassiter and Goodman's theory of the interpretation of gradable adjectives. In this account, MinAAs' thresholds are treated like any other gradable threshold, i.e. as an unbound variable whose meaning is inferred by the hearer through reasoning about the speaker's utterance choice and communicative intentions. This means that this theory predicts genuinely context-sensitive uses *and* interpretations of non-negated MinAAs to exist. More importantly, such context-sensitive uses *and* interpretations are claimed to arise via a coordination process between hearer and speaker that I have argued does not take place in the case of MinAAs.

Pragmatic thresholds

Whether vagueness and imprecision should be treated as one or two unrelated types of context-sensitivity is a matter of debate (see Solt 2015 for a recent overview of some of the arguments). In theories that assume shifting semantic thresholds for all GAs such as Lassiter and Goodman's, vagueness and imprecision are collapsed in one continuous category. Due to the type of world-knowledge priors that are generally associated with RelAs (which tend to be roughly normally distributed), and AAs (which tend to be skewed), the threshold posterior distributions resulting from Bayesian update are usually more peaked for AAs than RelAs. Therefore, in this theory the

distinction between vagueness and imprecision can be reduced to a difference in the uncertainty about the actual value of the lexical adjectival threshold. RelAs involve more uncertainty than AAs because their threshold posterior distribution has a bigger standard deviation than that of RelAs, thus making the value of relative thresholds more uncertain than that of absolute thresholds.[17]

Theories that treat vagueness as a semantic phenomenon and imprecision as a pragmatic one implicitly or explicitly assume a *pragmatic threshold* that regulates context-sensitivity arising from imprecision calculation. Although Lasersohn does not incorporate pragmatic thresholds in the formalization of the pragmatic halo, the notion of a pragmatic threshold is compatible with his proposal that the size of the pragmatic halo depends on *how much* deviation from the truth is tolerated in a given context. Other proposals, inspired in part in Lasersohn's seminal work on imprecision, overtly incorporate pragmatic thresholds as part of the formal representations of MaxAAs. For instance, van Rooij (2011), who adopts a delineation approach to model the relative/absolute distinction, proposes that *all* uses of an MaxAAs require setting the value of a contextual threshold of imprecision. Like most delineation approaches, van Rooij assumes that the variability in interpretation of RelAs is driven by Comparison Class variance (Klein 1980; Sassoon & Toledo 2011; Burnett 2014). MaxAAs' insensitivity to context can be straightforwardly explained if these predicates are systematically evaluated with respect to the maximal domain of individuals in the model. Since Comparison Class variability is eliminated, the criteria of application of these predicates remains constant across contexts and thus no standard-shift effects are predicted to arise with MaxAAs. In

17. It should also be pointed out that in this proposal the relative vs. absolute distinction is not a property of the predicates themselves or the scales used for their interpretation, but rather a result of the background knowledge that speakers and interpreters carry with them. Lassiter and Goodman welcome this as a desired outcome of their proposal, as it accounts for well-known counterexamples in which AAs seemingly give rise to relative-like interpretations depending on the Comparison Class used for interpretation. For instance, the standard of fullness is by convention much lower if the Comparison Class used for the interpretation of the adjective *full* are glasses of wine compared to cases in which the Comparison Class are glasses of beer. In Lassiter and Goodman's proposal, this Comparison Class based variability can be easily captured as differences in the background information that conforms the prior (although see McNally (2011) and Toledo & Sassoon (2011) for other proposed solutions to this and similar counterexamples).

this account, the threshold of MaxAAs is always endpoint-oriented. Context-sensitive interpretations of absolute standards involving non-endpoint-oriented thresholds result from changes in the *contextual standard of precision*. van Rooij formalizes this pragmatic type of context-sensitivity by assuming a set of models $\mathbb{M}$ that share both the domain of contexts C and individuals I, but differ in the standard of precision each of them imposes on the use of the predicate. In other words, each model in $\mathbb{M}$ is associated with a different valuation function V, such that for any two models M and M' in $\mathbb{M}$, it is the case that $V_M \neq V_{M'}$. The axis of variation across valuation functions is in the number of distinctions they make among the individuals in I with respect to a property P. For instance, a model M' makes more distinctions than a model M with respect to property P iff there exist two individuals such that M' distinguishes them with respect to P, whereas M does not, mapping them to the same degree of P-ness. In (42), this is modeled by means of the binary relation $\sim P$ (which could be paraphrased as *being as P as*) that establishes its two individual arguments are indistinguishable with respect to the property P.

(42) $\exists x, y \in I, M \vDash x \sim_P y$, but $M' \nvDash x \sim_P y$

When a model M' makes more distinctions than model M as in (42), M' is considered a *refinement* of M. In order for the distinctions made by a given model M to be preserved in finer-grained models M', the constraint in (43) must obtain. By using the subsethood relation, (43) ensures that any distinction made by a model M with respect to the valuation of the relation $>_P$ (*being P-er than*) will be preserved in any model M' that is at least as fine-grained M. Thus, in this system, context-sensitive uses of AAs are not captured as adjectival threshold shifts, but rather as shifts in the contextual threshold of imprecision.

(43) M' is a **refinement** of M with respect to predicate P iff $V_M(>_P) \subseteq V_{M'}(>_P)$

In a similar vein, and again building on insights from Lasersohn (1999), but also Kennedy (2007) and Potts (2008), among others, Leffel *et al.* (2016) propose that resolving the meaning of a MaxAA involves fixing two thresholds: the standard semantic threshold, which is assumed to

be insensitive to contextual variability, and a pragmatic threshold that determines what deviations from the maximal scalar endpoint are allowed in the context (44).

(44) a. The pragmatic threshold ${\theta_A}^{imp}$ for a MaxAA GA A is the largest deviation from max (D_A) that is pragmatically ignorable in a context.

b. The set of pragmatic alternatives to A_{pos} is the set
$Alt(A_{pos}) = \{\lambda x.\mu_A(x) \succeq (\mathbf{max}_A\text{-}{\theta_A}^{imp}) \mid {\theta_A}^{imp}$ is a pragmatic threshold$\}$

Therefore, for theories that treat MaxAAs as only being subject to imprecision calculation, the meaning of these predicates can be thought of as involving two types of thresholds: a semantic adjectival threshold, and a pragmatic threshold that modulates the degree of precision assumed for the interpretation of the adjective.

2.2.4 Interim Conclusion

The two types of approaches discussed in this section (i.e., theories that assume shiftable absolute thresholds and theories that assume fixed absolute thresholds) make the correct predictions with respect to the range of possible interpretations available for RelAs and AAs. Unfortunately, discriminating between these two families of approaches proofs difficult because there do not exist good test to diagnose whether a threshold regulates semantic or pragmatic phenomena. For this reason, I argue that a shift in methodology might turn out to be a productive way to shed some new light on this question. Specifically, I propose that online processing, and in particular, methodologies that have been shown to be sensitive to the interactions between context and interpretation are a good testing arena to look at the kind of issues that constitute the focus of this dissertation. In the following section, I introduce one of such methodologies, i.e. the Visual World eye-tracking paradigm, and review studies that have used this methodology to address the question of how contextual information is integrated during online processing of prenominal adjectives.

2.3 Context-sensitivity during online processing of Gradable Adjectives

There exists ample evidence that listeners process linguistic input incrementally (Crain & Steedman 1985; Altmann & Steedman 1988; Eberhard *et al.* 1995, among many others), and that pragmatic information pertaining to different sources is quickly integrated during online processing (Hanna *et al.* 2003; Hanna & Tanenhaus 2004; Grodner & Sedivy 2011). In particular, the Visual World eye-tracking paradigm has proven useful in investigating questions about the influence of context during the online resolution of temporarily ambiguous referents (see Huettig *et al.* 2011 for a recent overview of applications of the VW paradigm to the study of language processing). In this methodology, participants' eye-movements are recorded while looking at visual displays as they hear utterances. Eye-movements are a particularly good measure of language processing in reference-resolution tasks because eye-fixations reflect with millisecond granularity what objects in the visual context are being considered as potential referents of the linguistic input (Cooper 1974; Eberhard *et al.* 1995; Tanenhaus *et al.* 1995; Pyykkönen-Klauck & Crocker 2016). In general, experimental manipulations vary features of the visual context, as well as the properties of the linguistic input. Based on the eye-movement patterns that emerge as a function of these manipulations, researchers make inferences regarding the ways in which contextual information is integrated during linguistic online processing. For instance, in seminal work on this paradigm, Tanenhaus *et al.* (1995) showed that contextual information, introduced by the manipulation of the visual display, was immediately adopted by the participants to guide their online parsing decisions.

2.3.1 VW studies of prenominal adjectives

In this section, I focus on experimental results that make use of the Visual World paradigm to investigate the influence of context in reference resolution of definite descriptions involving prenominal adjectival modification. Within Visual World studies, *Referential Effects of Contrast* (henceforth RECs) constitute a hallmark of this rapid online integration of contextual information during online processing. The effect was first reported by Sedivy et al. (1999) in a study investigating how prop-

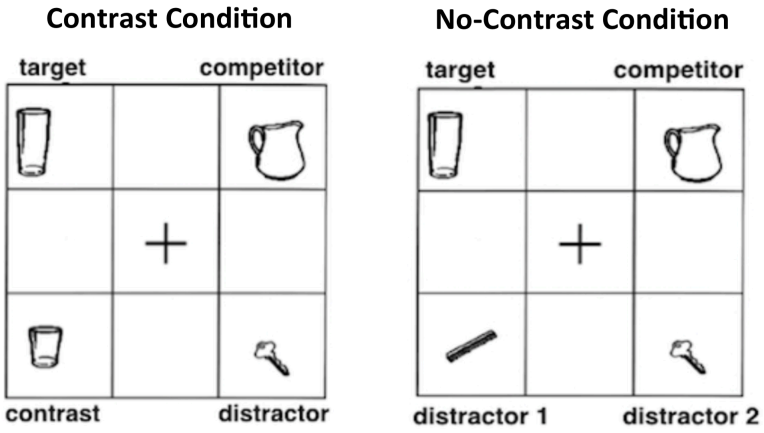


Figure 2.6: Trial example of the visual stimuli used in Sedivy *et al.* (1999) study.

erties of the visual context influence the processing of NPs containing an attributive prenominal adjective like *tall*. In particular, Sedivy *et al.* (1999) investigate how the presence of a contrast set (or Comparison Class) in the visual display influences the processing of sentences containing relative adjectives like *tall*. In the experiment, participants heard instructions such us *Pick up the tall glass* while looking at displays of four objects. Two conditions were tested: the Contrast condition supported a contrasting interpretation of the adjective by including, alongside the target object (e.g., a tall glass), a contrast object that could be described by the noun but not the adjective in the instruction (e.g., a short glass). In the second condition, the No-Contrast condition, the contrasting object was substituted with a distractor, i.e., an object that could not be described either by the head noun or the modifier in the instruction. Finally, both conditions contained a competitor that presented a higher degree of the property in the instruction when compared to the target, but was judged to be better described by an unmodified noun (e.g., a pitcher that was tall*er* than the glass, but was itself not tall for a pitcher, se Figure 2.6).

The main finding of the experiment was that participant's fixations converged on the target faster in the Contrast condition than they did in the No-Contrast condition. Crucially, as seen in Figure 2.7 in the Contrast condition participants zoomed into the target object during the adjective

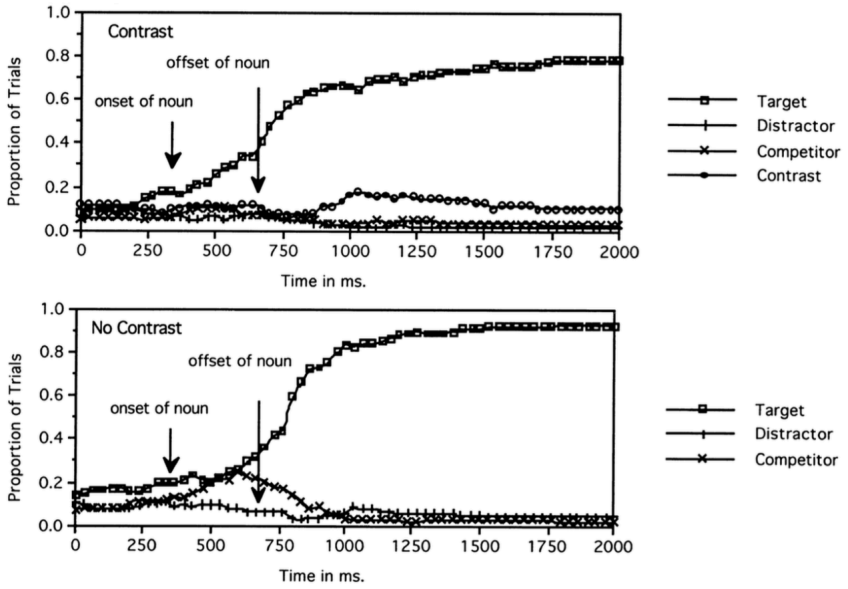


Figure 2.7: Results from Experiment 2 as reported by Sedivy *et al.* (1999).

window at a point in which the head noun had not yet been processed. This early identification of the target object was in part due to the fact that participants stopped fixating on the competitor object earlier in the Contrast condition compared to the No-Contrast condition. Remarkably, participants were able to identify the target at a point in which the linguistic instruction was still compatible with either the target or the competitor, suggesting that the information about the contrasting object was used very quickly.

Finally, in a recent study modeled after Sedivy's (1999) study, Leffel *et al.* (2016) present results from a Visual World eye-tracking study testing RelAs and imprecise interpretations of MaxAAs. In a nutshell, their results showed that RelAs gave rise to RECs when the target object was a better exemplar of the adjectival property than the competitor (e.g. the target was a taller object than the competitor, and the auditory instruction was *Click on the tall NP*), whereas the effect disappeared when the competitor instantiated the adjectival property to a higher degree compared to the target object (e.g. the competitor was a taller object than the target, and the auditory instruction was *Click on the tall NP*). MaxAAs showed the opposite pattern of effects: the REC was robust when the

competitor was a good exemplar of the adjectival property, whereas the condition in which the target object was a better exemplar than the competitor only gave rise to a mild REC. Based on this, Leffel *et al.* argue that the puzzling behavior of MaxAAs can be explained if this class of adjectives is sensitive to pragmatic reasoning about imprecision, a type of reasoning that is only triggered when the context provides support for an imprecise interpretation of the adjective. I will have more to say about the results of this study in Chapter 3.

2.3.2 The role of informativity in RECs

Sedivy and colleagues (1999) point out that there are at least two interpretations compatible with their findings. The first one is that the REC observed for RelAs is driven by the use of the restrictive prenominal modifier, a use that triggers pragmatic reasoning about a set of referents that contrast with respect to the degree to which they possess the property denoted by the modifier (Altmann & Steedman 1988; Tanenhaus *et al.* 1995). If hearers assume that the speaker is observing the Gricean Maxim of Quantity, uttering the prenominal modifier would be maximally informative only in those cases in which the modifier is used to disambiguate between two possible referents. This view predicts that effects of referential contrast should arise with any prenominal modifier, as long as the context supports a restrictive interpretation.

A second option is that the effect of contrast is driven by aspects of the lexical semantics of RelAs. As discussed in §2.2.3, the threshold of RelAs must be resolved via context. Therefore, participants could be integrating the visual contextual clues as part of the lexical processing of the adjective itself. Under this view, the slower visual identification of the target object observed in the No-Contrast condition would reflect the difficulty of processing a RelA in the absence of appropriate contextual support. This account predicts that adjectives that do not have context-sensitive thresholds, should not be sensitive to the presence or absence of a contrasting object in the visual display. However, this prediction does not seem to be fully borne out. In a subsequent eye-tracking study that replicated the same method and design, Sedivy (2004) tested color and

material adjectives such as *wooden* or *plastic*. Since neither of these two adjective types have context-sensitive thresholds, no REC was predicted to arise. As expected, color adjectives did not reveal any significant difference between the Contrast and the No-Contrast condition. However, material adjectives did give rise to a REC such that trials containing a contrasting object showed facilitated target identification. This latter finding is somewhat surprising if the RECs are lexically driven. For this reason, Sedivy (2004) rejected a purely lexical account of RECs, and argued that listeners' pragmatic reasoning about utterance informativity must be responsible for these effects.

Informativity-based accounts of referential contrast might also be able to explain the absence of RECs for color adjectives reported in Sedivy (2004). In a follow-up eye-tracking study investigating color adjectives, Sedivy (2003) finds that the presence of a contrastive object only facilitated the identification of the target in those cases in which the color was predictable from the head noun, e.g., when the target was a yellow banana and the contrastive object was a brown banana. Sedivy argues that overt uses of predictable color adjectives must be taken to be highly informative by participants, as otherwise the speaker would be overspecifying by using a modifier that is redundant given the noun. This account is also compatible with the absence of contrast effects for color adjectives found in Sedivy 2004, where all color adjectives were unpredictable from the head-noun. Finally, a third logical possibility is that certain types of adjectives, like color adjectives, only show informativity-based effects of contrast while other types of adjectives, like RelAs, show a combination of both. This is precisely one of the research questions that the experiments reported in Chapter 3 investigate.

2.4 Conclusion

This chapter has presented an overview of the semantic and pragmatic properties of GAs, as well as the different theories that have been proposed in order to account for their similarities and differences. In particular, I have focused on the fact that RelAs and AAs display different degrees of context-sensitivity, with RelAs showing context-sensitive uses across the board; MaxAAs display-

ing a much more reduced context-sensitivity limited to how much deviation is allowed in the context from the endpoint-oriented interpretation; and MinAAs arguably not giving rise to context-sensitive interpretations altogether. An open question is whether the source of the context-sensitivity presented by RelAs and MaxAAs stems from the same source, or whether MaxAAs only show context-sensitive uses that involve pragmatic reasoning about imprecision. A second question is whether MinAAs should be modeled as context-sensitive predicates in the first place. Settling these matterns on the base of truth-conditional judgements has been proven a difficult a task. For this reason, I have proposed that experimental methodologies, such as the Visual World eye-tracking paradigm (§2.3)—which has been shown to efficiently track the integration of contextual information during online processing—can potentially provide new and useful data to better inform our theories of gradability in the adjectival domain. In Chapter 3, I present a sequence of two experiments, a Visual World eye-tracking study and an offline judgement study, that have been designed precisely with the objective of addressing these questions.

CHAPTER 3

PROCESSING GRADABLE ADJECTIVES IN CONTEXT

3.1 Introduction

The theories described in Chapter 2 can account for the different patterns of context-sensitivity associated with RelAs and AAs. However, adjudicating among the existing proposals on purely theoretical grounds is difficult, as both families of theories derive the correct meanings for the two classes of GAs. In this chapter, I report results from a Visual World eye-tracking experiment (Experiment 1, §3.2) that investigates how context influences reference resolution during online processing of GAs. The aim of the experiment is to determine whether contextual information is integrated in comparable ways during online processing of RelAs and AAs. A positive answer to this question would constitute support for theories that assume *shifting absolute thresholds* (§2.2.3), whereas the opposite finding would suggest that contextual information is recruited differently during the processing of these two adjective types, thus lending support to theories that argue in favor of *fixed absolute thresholds* (§2.2.3). The second experiment presented in this chapter (Experiment 2, §3.3) is designed as a follow up to Experiment 1 and has the goal of quantifying the overspecification penalty associated with redundant uses of the four classes of adjectives tested in Experiment 1, with the objective of controlling for potential confounds in the interpretation of the eye-tracking results.

3.2 Experiment 1: Variable RECs across Adjective Classes

The current study builds upon Sedivy *et al.*'s findings. In particular, I focus on the question of whether listeners' rapid sensitivity to context is driven by informativity-based pragmatic reasoning, lexical-semantic processing of a Comparison Class, or a mixture of both. A better understanding of the contributions of these different factors to the processing of GAs will also allow us to more precisely locate the source of the context-sensitivity of RelAs and AAs. With these goals in mind,

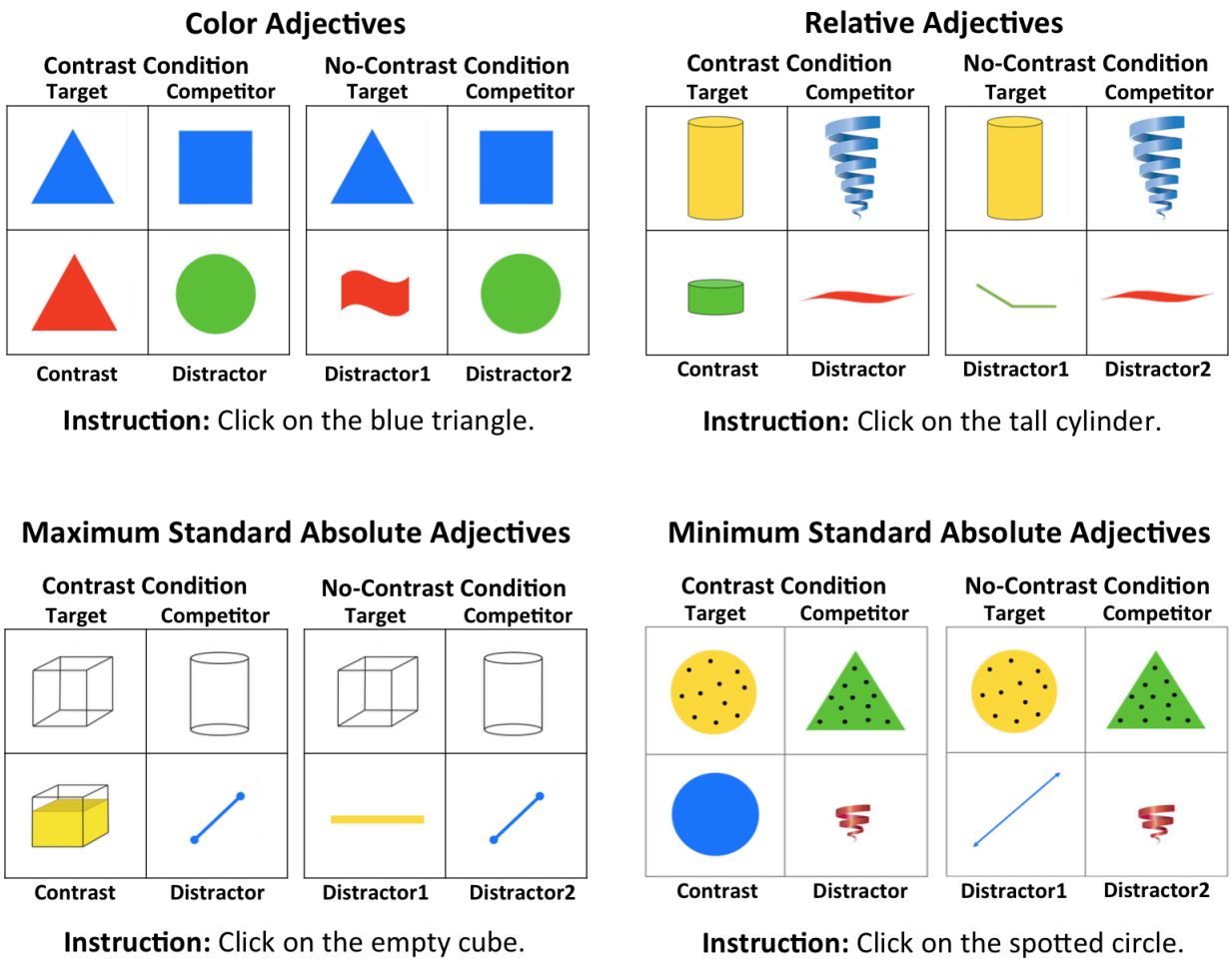


Figure 3.1: Item Examples.

the present study replicates Sedivy *et al.*'s design and extends it to AAs. However, I depart from their design in that I do not use real world objects, but rather geometric shapes. I do so in order to control for head-noun effects that might result from world knowledge. Since geometric shapes are idealized mathematical objects, I assume that they should give rise to less variability based on world knowledge.

Four types of adjectives were tested in one experiment: RelAs (n=9), MaxAAs (n=4), MinAAs (n=6) and Color Adjectivess (n=4, henceforth ColAs. See Table 3.1 for an exhaustive list of all the adjective-noun pairs tested in the experiment). Following Sedivy *et al.* (1999), two critical kinds of

visual displays were tested (see Figure 3.1). In the Contrast condition, the visual display contains: 1) a TARGET object (e.g., a tall cylinder) that participants are requested to click on; 2) a COMPETITOR that shares the target property but presents a different shape (e.g., a tall spiral);[1] 3) a CONTRAST object that belongs to the same Comparison Class as the target, but could not be described by the adjective in the instruction (e.g., a short cylinder); and 4) a DISTRACTOR object that could not be described by the adjective in the instruction, nor does it belong to the same comparison class (e.g., a wavy line). The No-contrast condition was created by substituting the contrasting object with a second distractor. With the exception of ColA trials, none of the shapes in the visual array shared color. Finally, all the target and competitor pictures used in AAs trials were consistent with precise interpretations, as they supported an interpretation of the predicate that was compatible with an end-point oriented standard.[2]

Sixty experimental items (20 containing RelAs, 10 containing MaxAAs, 10 containing MinAAs and 20 containing ColAs) were constructed.[3] The adjective-noun pairs tested in the experiment are presented in Table 3.1.[4] Conditions were distributed in two lists using a Latin Square design. Both the order of the trials within each list and the position of the four pictures within each trial were randomized.

Each list was complemented with 60 filler trials. All adjectives used in filler trials were ColAs (*red*, *green*, *yellow* and *blue*), and pictures always consisted of 2D shapes with plain colors.

---

1. The competitor objects used in the current study differ from Sedivy et al.'s design in that they all presented the same degree of the adjectival property as the target object, and they could all be felicitously described by the adjective in the auditory instruction.

2. See results from the norming studies below.

3. The full list of the experimental items used in Experiment 1 can be found in the following link:
http://lucian.uchicago.edu/blogs/lpl/exp1_aparicio_dissertation/

4. The bigram frequency of each of the adjective-noun pairs involving a gradable adjective in Table 3.1 was calculated using the Brown corpus. Most of the bigrams were unattested. Only three combinations were attested: *straight line* occurred 12 times, *thin line* occurred 2 times. Finally, *long line* was attested once.

RelA	Noun
long	line
short	spiral/cylinder/line
tall	spiral/cylinder
big	square/triangle
small	square/triangle
thick	line
thin	line
wide	rectangle/oval
narrow	rectangle/oval

MaxAA	Noun
closed	circle
flat	circle/triangle
full	cube/cylinder
empty	cube/cylinder

MinAA	Noun
bent	line
bumpy	square/triangle
curved	line
open	circle
spotted	square/circle
striped	square/triangle

ColA	Noun
blue	circle/triangle/square/line
red	circle/triangle/square/line/oval
green	circle/triangle/square/rectangle
yellow	circle/triangle/square/line/rectangle

Table 3.1: Adjective-Noun pairs tested in the experiment.

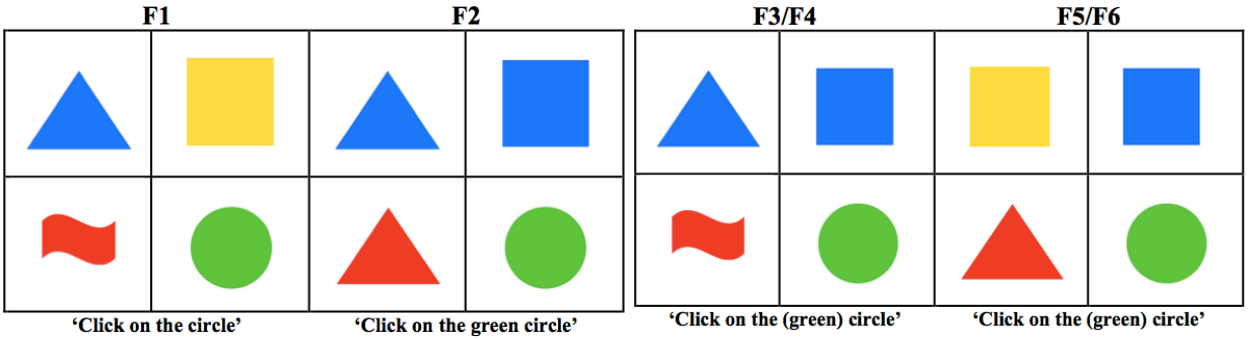


Figure 3.2: Fillers

Six different types of fillers (10 trials per type) were constructed (see Figure 3.2). In the first type (F1), none of the figures shares shape or color and the instruction does not contain a modifier. In the second type of filler (F2), the visual display is equivalent to the Contrast condition in the color-adjective trials. However, these filler trials differ from the Contrast condition in that the auditory instruction targets the distractor. In the third type of filler (F3), none of the objects share shape, although two of the pictures share color. The instruction contains a modifier but it does not target any of the two shapes that share color. The fourth type of filler (F4) only differs from F3 in that the instruction does not include a modifier. In the fifth type of filler (F5), none of the figures in the visual array shares color. However, two of the shapes belong to the same comparison class. The instruction contains a modifier and targets one of the two pictures that does not share shape with any of the other pictures in the visual array. Finally, the sixth type of filler (F6) is like F5, except that the instruction does not make use of a color adjective.

3.2.1 Predictions

Informativity

I now turn to the predictions made by the current experimental design regarding quantity and manner-based reasoning about informativity associated with the use of a prenominal modifier. A naive pragmatic view assumes that such reasoning should apply to all instances of prenominal modification. If this is the case, it is expected that the four adjective classes tested in the current experiment should give rise to an informativity-based REC. Furthermore, if informativity considerations are the only driver of RECs, it is also predicted that there should not be significant differences in the timing, magnitude and/or qualitative properties of the hypothesized RECs.

If, on the other hand, both general pragmatic reasoning and lexical context-sensitivity jointly contribute to RECs, the following is predicted: ColAs should provide us with a baseline for the

informativity-based REC linked to the use of restrictive modification.[5] Furthermore, based on Sedivy *et al.*'s (1999) previous results on RelAs, it is expected that RelAs should also show a REC. Crucially, given that RelAs, unlike ColAs, are lexically context-sensitive, the REC displayed by RelAs should be qualitatively different from the REC displayed by ColAs, since the former would not only be shaped by informativity-based pragmatic reasoning, but also by lexical processing.

Predictions made by theories assuming variable absolute thresholds

I now focus on the predictions that theories assuming variable absolute thresholds make for the current experiment. In Chapter 2, I presented Lassiter & Goodman's (2013) account as representative of such position. Lassiter & Goodman put forth a RSA model that can derive the available range of interpretation of GAs. When spelling out the predictions that this model makes with respect to online processing, one should proceed with caution, since the relationship between RSA models and online processing remains a poorly understood question. To put it differently, the logical space of possible linking hypothesis relating specific features of RSA models to online measures of linguistic processing has barely been explored. For this reason, the assumptions I will be making about this relationship will be minimal, namely that such relation exists, with the hope that the results obtained in the current experiment will constitute a first step towards bridging the gap between these two domains.

Because of its elegant simplicity, the model proposed by Lassiter & Goodman could be taken to constitute the null hypothesis, i.e., that there are no significant differences across GAs with respect to the way in which contextual information is recruited in order to arrive at an interpretation. Based on this, the simplest prediction made by this account regarding Experiment 1 is that all the GAs tested in this experiment should show comparable patterns of integration of contextual information

5. Although ColAs manifest various types of context-dependence (see, e.g., Kennedy & McNally 2010), they have not been argued to show the same kind of Comparison Class-based context dependence that constitutes the focus of this study.

from the visual context, as part of the process of resolving the threshold variable of the GA. The prediction regarding ColAs is nevertheless different. Unlike GAs, the lexical representation of ColAs does not contain a threshold variable. Therefore, the patterns of integration of contextual information observed for ColAs should differ from those displayed by GAs.

Predictions made by theories assuming fixed absolute thresholds

Theories that assume fixed absolute thresholds predict richer patterns of integration of contextual information among GAs. First, since both RelAs and MaxAAs are context-sensitive, it is predicted that they should both give rise to RECs. However, there should be non-trivial qualitative differences between the RECs of RelAs and MaxAAs, since in the former case information from the visual context is claimed to be used to fix the value of a semantic threshold, whereas in the latter case this information is used to fix the value of the pragmatic precision threshold. Second, since I have claimed that the interpretation of MinAAs is not sensitive to context, MinAAs should pattern like ColAs in displaying an informativity-driven REC, assuming that this kind of pragmatic reasoning applies to all instances of prenominal modification.

3.2.2 Methods

Materials

Visual Stimuli

Pictures used in experimental trials as targets, contrasts and competitors (a total of 108) were normed in a series of three description-picture matching studies on Mechanical Turk. The purpose of the norming studies was to standardize the interpretational preferences of the visual stimuli within and across adjective types. More specifically, the norming studies ensured that all target and competitor objects were recognized to satisfy the relevant adjectival property, whereas contrast objects (used in the Contrast condition) were recognized to NOT instantiate the relevant adjectival

property. In addition, 81 more images were used as distractors. Whenever possible, distractors were drawn from the pool of objects that had been used as target, competitor or contrast in other trials.

For the first norming study, **Norming Study 1**, a total of 162 scales (1,134 pictures) were created using the gradable adjectives in Table 3.1.[6] Each scale consisted of a 7-point continuum along the property denoted by the adjective. In all scales, the degree of the property was increased progressively from left to right (see Figure 3.3 for a RelA scale example, Figure 3.4 for a MaxAA scale example and Figure 3.5 for a MinAA scale example.). The same four colors (red, blue, green and yellow) were used in all the scales. The objective of this norming study was to get a sense of the range of the scale for each adjective type in order to select the best scale-points to be used in the eye-tracking experiment.

Description: *This is a tall cylinder.*

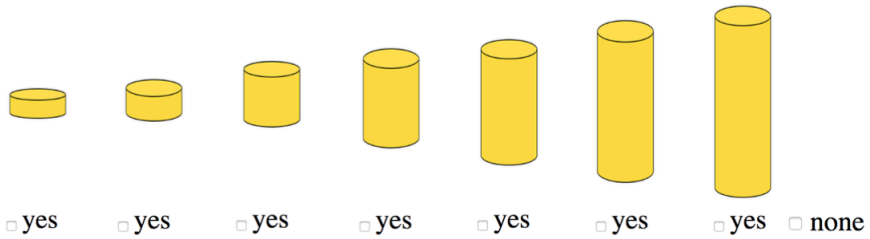


Figure 3.3: Norming study 1, RelA trial example.

Description: *This is an empty cube.*

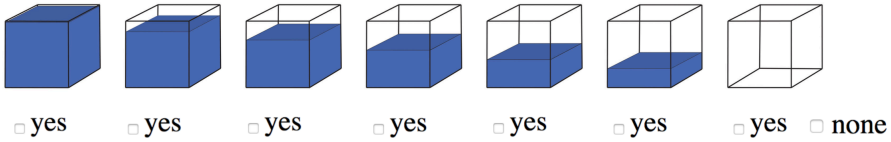


Figure 3.4: Norming study 1, MaxAA trial example.

The 162 scales were tested in three different Mechanical Turk experiments. Participants were presented with a series of 7-point scales, each accompanied with a statement of the form *This is*

6. Visual stimuli used in ColA trials were a subset of the stimuli used in RelA and AA trials and were not normed. See information about filler trials below for further information about ColA visual stimuli.

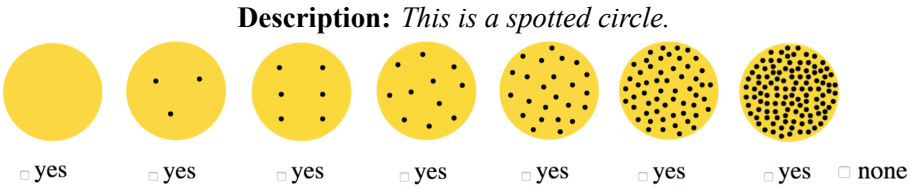


Figure 3.5: Norming study 1, MinAA trial example.

a tall cylinder (see Figures 3.3-3.5 for trial examples). The description made use of the indefinite article, instead of the definite article *the*, in order to avoid the existential presupposition carried by the definite, since it could have biased participants towards the existence of the object described in the statement. Participants were instructed to check the *yes* box under those pictures that they thought matched the description. If they considered that none of the pictures fitted the description, they were instructed to click on the box *none*. The scales were pseudo-randomized in three lists and participants were randomly assigned to each of these three lists. Participants were 108 native speakers of American English between the ages of 18-35 (47 females, mean age 27.1).

Two criteria were followed in order to discard problematic scales. First, I obtained the proportion of *none*-responses for each scale. An arbitrary threshold was set at a low proportion of 10% such that scales that had obtained more than 10% of *none*-responses were rejected. Second, I discarded scales for which the proportions of *yes*-responses did not follow the prototypical patterns expected for each adjective type. For RelAs, the proportions were expected to increase progressively along the scale. For MinAAs, I expected to find high proportions of *yes*-responses starting at the second point of the scale and to remain high all the way to end of the scale. Finally, for MaxAAs, a low proportion of *yes*-responses was expected with the exception of the last item in the scale (i.e. scale point 7), which should present a high proportion of *yes*-responses, since it is the only point that matches the description. Following these two criteria, a total of 10 scales were eliminated. Figure 3.6 shows the plotted proportions of *yes*-responses for each adjective type after eliminating the problematic scales. As can be observed in the plots, overall each adjective type

presents the prototypical profile of *yes*-responses.

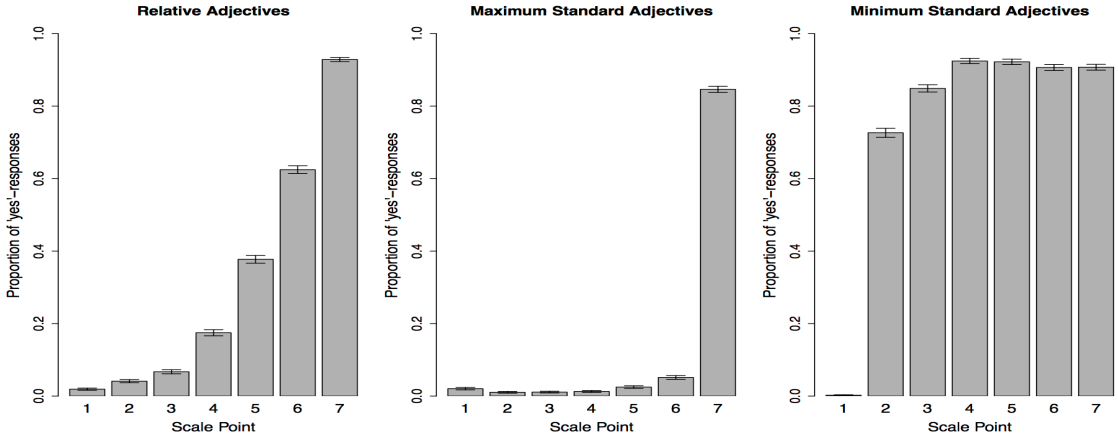


Figure 3.6: Proportions of yes-responses for each adjective type

In order to get a sense of where the threshold falls for each adjective type, the mean of the first scale-point at which each subject started answering *yes* was obtained. The mean values for each adjective type are plotted in Figure 3.7.

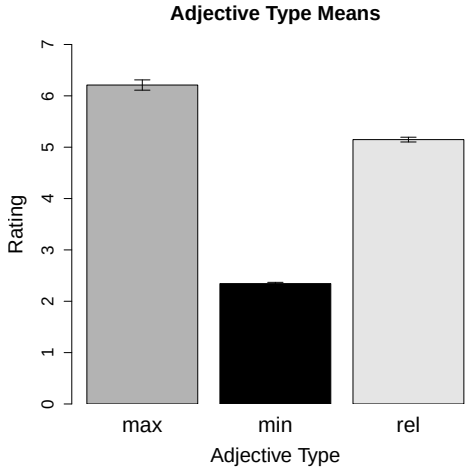


Figure 3.7: Means of scale-point at which participants started answering *yes* for each adjective type

As expected, MaxAAs present the higher mean (above 6 in the 7-point scale), whereas MinAAs present the lowest means (around 2 in the 7-point scale). For RelAs, the mean is slightly higher than 5, suggesting that for many of the scales used in the three experiments, participants thought the threshold was not in the mid point of the scale, but rather toward the end. A one-way ANOVA

showed that the means are significantly different, revealing a main effect of adjective type ($F(2, 68) = 253.79$ $p < .05$).

Based on these results, two points of the scale were chosen to be tested in the actual experiment. The choice of scale-points was done such that, for each adjective type, one of the scale-points would correspond to objects that could be clearly judged as having the property (i.e. objects that would later be used as targets and competitors), and objects that clearly did not instantiate the adjective property (i.e. objects that would be later used as contrasts). Based on the proportion of *yes*-responses in Figure 3.6, and the values of the thresholds in Figure 3.7, points 3 and 7 were chosen for RelAs and MaxAAs, whereas for MinAAs the scale-points selected were 1 and 4. The remaining norming studies have the goal of ensuring that the judgements about each of these scale-points are stable when these pictures are judged in pairs and in isolation.

In **Norming Study 2** the four scale-points selected in the first norming study were judged by participants in isolation. This was done in order to ensure that these pictures gave rise to similar judgements as those observed in Norming Study 1, even when there were no other members of the same comparison class present. In the current eye-tracking experiment, this is the case for all competitor objects and for all target objects in the No-Contrast condition.[7]

The four scale-points were distributed in three different Mechanical Turk experiments. A total of 104 two-condition items were created using the scale-points pairs 1-4 and 3-7. For each experiment, conditions were distributed in two lists following a Latin Square Design. In each trial, participants saw each of the scale points in isolation. All pictures appeared inside a black frame that was intended to provide a spacial point of reference when performing judgements involving dimensional adjectives like *tall/short* or *big/small*. As in Norming Study 1, each picture was paired with a statement of the form *This is a bumpy square* (see Figure 3.8 for a trial example belonging to each

7. For completeness, pictures corresponding to scale-points 1 and 3, which were used as contrasts, were also included in Norming Study 2, even though these scale-points never appear in isolation in either of the two experimental conditions tested in the eye-tracking study.

adjective type tested). Three possible answers were provided: *yes*, *no* and *neither*. Participants were 102 native speakers of English between the ages of 18-35 (54 females, M=27.39).

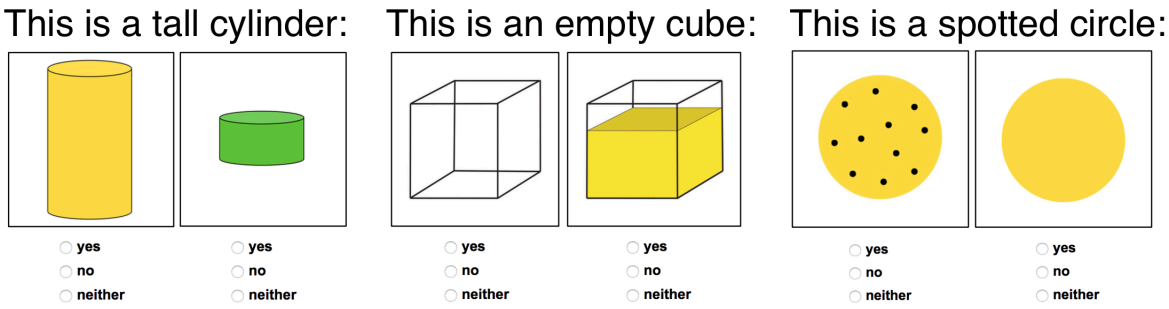


Figure 3.8: Norming Study 2 trial examples for each of the three adjective types (RelAs, MaxAAs and MinAAs) tested.

The proportion of *yes*-responses for each adjective type and scale-point tested are presented in Figure 3.9. As expected, scale-points 4 and 7 received a high proportion of *yes*-responses (above 90% across the board), confirming that these scale-points constitute a good fit with the description used in each trial. Scale-points 1 and 3 received a very low proportion of *yes*-responses (below 20% for RelAAs and below 10% for MaxAAs and MinAAs), validating that these objects could not be appropriately described by the adjective used in the description. Results from three paired t-tests confirmed that the means for each scale-point within each adjective type were significantly different (RelAs: $t(54) = -25.91$, $p < 0.001$); MaxAAS: ($t(21) = -50.68$, $p < 0.001$); MinAAs: ($t(26) = -21.46$, $p < 0.001$). To sum up, results from Norming Study 2 are robust enough to give us sufficient confidence about the quality of the tested visual stimuli when judged in isolation.

The third Norming Study, **Norming Study 3**, was conducted once the experimental lists had been created. The purpose of this study was to test those pairs of images used as the contrast set in Contrast Condition trials. Given that the images used in the experimental trial displays never shared color, each contrast set was drawn from two different scales that differed only in color (see Figure 3.10 for trial examples).

Forty contrast sets were tested in a Mechanical Turk experiment. For each contrast set, two orderings were created in which the horizontal position of the pictures with respect to each other

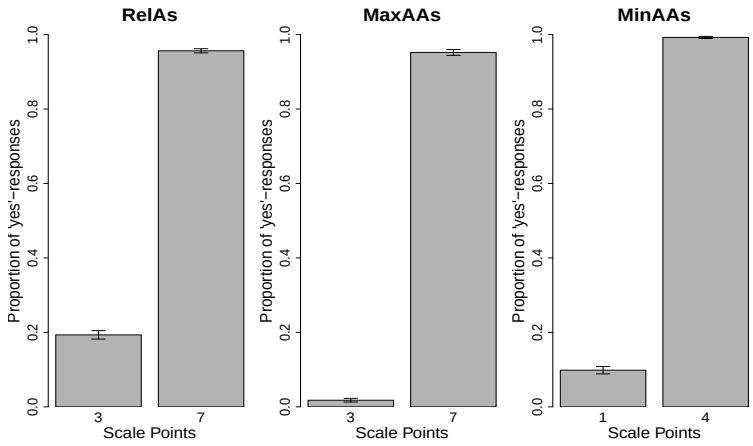


Figure 3.9: Proportions of '*yes*'-responses for each scale point and adjective type tested in Norming Study 2.

was manipulated. This was done in order to not bias participants to expect one particular scale-point to always appear to the left or to the right of the visual display. The two orderings were distributed over two lists using a Latin Square design. As in Norming Study 2, each pair of images appeared inside a black frame. Each contrast set was paired with a statement of the form *This is a bumpy square* and three possible answers (*yes*, *no* and *neither*) were provided. Participants were 34 native speakers of English between the ages 21-35 (18 females, M=28.54).

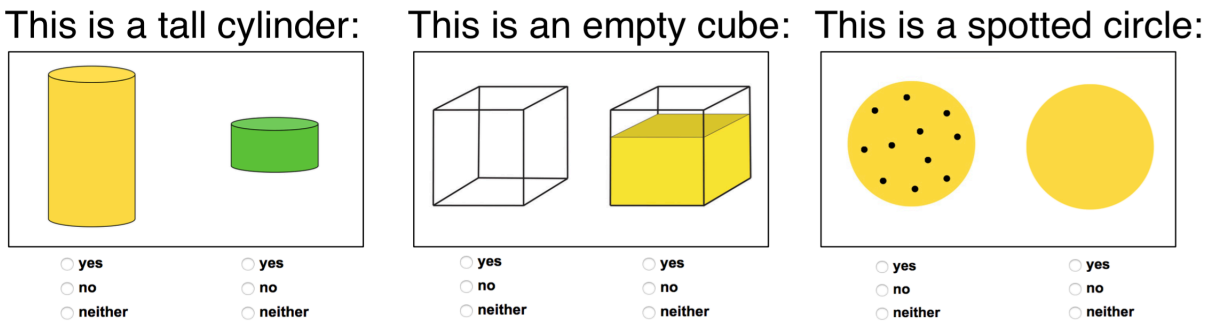


Figure 3.10: Norming Study 3 trial examples for each of the three adjective types (RelAs, MaxAAs and MinAAs) tested.

Proportions of *yes*-responses for each adjective type can be found in Figure 3.11. As can be observed in the plot, scale-points 4 and 7 again received a high proportion of *yes*-responses, above

90%, whereas scale-points 1 and 3 received very low proportions of *yes*-responses, below 10% for all adjective types. Three paired t-tests confirmed that the means for each scale point tested for each adjective type were indeed significantly different (RelAs: $t(19) = -9.03$, $p < 0.001$); MaxAAs: $t(9) = -32.99$, $p < 0.001$); MinAAs: $t(9) = -104.33$, $p < 0.001$)).

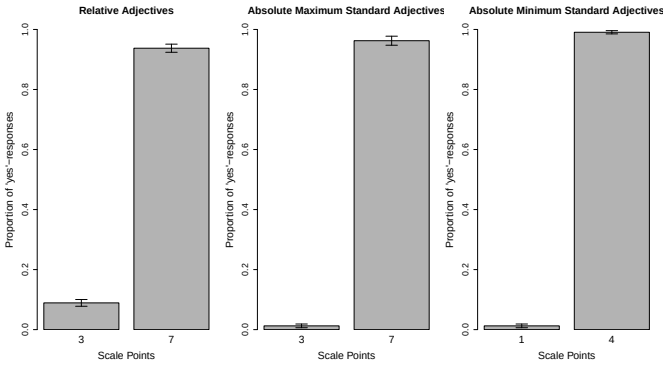


Figure 3.11: Proportions of *yes*-responses, Norming Study 3

Taken together, the results of the three norming studies presented above provide us with a good measure of the interpretation preferences of the tested stimuli and the properties of the visual stimuli used in the eye-tracking experiment.

Auditory Stimuli

Auditory stimuli were recorded in a sound booth by a female native speaker of English. For each recording, the onsets and offset of the adjective were measured in order to determine the mean duration of the three groups of adjectives tested. For RelAs, the mean duration was 469 ms (SD = 74.17). For MaxAAs the adjective mean duration was 442 ms (SD = 53.12). Mean duration for MinAAs of the adjective for all trials was 503 ms (SD = 76.09). Finally, for ColAs the mean duration was 466 ms (SD = 46.09). The general mean for all adjective types is 469.5 ms. A linear regression revealed that adjective type was a significant predictor of adjective length ($F(3,116) = 3.23$, $p < 0.03$). However, the differences turned out to be uniquely driven by the longer duration of MinAAs when compared to ColAs (RelAs: $\beta = 2.65$, $SE = 14.098$, $p > 0.8$; MaxAAs: $\beta = -24.15$, $SE = 17.266$, $p > 0.1$; MinAAs: $\beta = 37.25$, $SE = 17.266$, $p < 0.05$). None of the adjectives bore

pitch accent or rising tone.

Apparatus

Eye movements were recorded with a Tobii T60 Eye-tracker sampling at 60 Hz. Viewing was binocular and both eyes were tracked, although analyses were performed on data belonging to the right eye exclusively.

Procedure

Participants saw a visual display with four pictures. Their eye movements were tracked while listening to instructions such as *Click on the tall cylinder*. Participants were instructed to click on the picture that they thought fitted the description in the auditory instruction best. Only clicks that took place after the offset of the auditory instruction triggered the next trial. There was a 2-second long preview window between the onset of the visual display and the onset of the auditory instruction. Before each trial, a fixation cross appeared in the middle of the screen. A red box framing the cross appeared when participants fixated on it. Participants were instructed to click on the cross when the red box appeared in order to proceed to the next trial. This was done so that eye movements to the four objects could be measured from a default position that was equidistant to the four pictures in the display. At the beginning of the experiment, participants had four practice trials to help them become familiar with the task.

Participants

Participants were fifty-one undergraduate and graduate students at the University of Chicago (34 females, M = 20.7, range 18-34). All participants were native speakers of American English. Undergraduate students did the experiment to fulfill a research awareness requirement for a linguistics course. Graduate students were paid $10. All participants had normal or corrected to normal vision. Subjects were excluded from data analysis if they met at least one of the following two criteria: 1)

track loss for a given subject was higher than 40%; and 2) before the head noun became available, a subject did minimal scanning of any part of the display (i.e., when the aggregated proportion of fixations to the four pictures in the display was 10% of the total recorded fixations, probably because the subject was only fixating on the fixation cross in the center of the screen). The latter criterion intends to exclude participants who were passively waiting for the head noun information before processing the instruction. The application of these two criteria resulted in the exclusion of 11 subjects. The results reported in the following section correspond to data from 40 participants between the ages of 18-34 (26 females, M = 20.57).

3.2.3 Results

Analyses were run on the aggregated proportion of fixations. For each trial, the adjective window was offset by 200 ms from the onset of the adjective to adjust for the time required to plan and implement an eye-movement. One MinAAs-noun combination was removed from data analysis because the stimuli was found to not be recognized as appropriately representing the adjectival property within 1200 ms of the adjective onset.

Analysis I: Target vs. Competitor Disambiguation

In this section, I focus on the timing of target vs. competitor disambiguation in each of the two conditions tested. The goal of this comparison is to determine whether the presence of the contrasting object facilitated target identification in the Contrast condition when compared to the No-Contrast condition.

For each adjective type, analyses were performed on two consecutive windows (W1 and W2) of 150 ms, such that the right boundary of W2 coincided with the mean onset of the head noun in the auditory instruction. A third window (W3) of 150 ms starting at the onset of the head noun was also analyzed. W1 and W2 contain fixations reflecting the processing of the adjective, whereas W3 contains fixations reflecting the processing of the head noun (see Table 3.2). Figure 3.12 contains

the proportions of fixations to each of the four objects in the visual display for each condition. Eye fixations to the target and the competitor objects were analyzed.

Adjective Type	W1	W2	W3
ColAs	355-516 ms	517-667 ms	668-818 ms
RelAs	369-519 ms	520-670 ms	671-821 ms
MaxAAs	342-492 ms	493-643 ms	644-794 ms
MinAAs	403-553 ms	554-704 ms	705-855 ms

Table 3.2: Time windows defined for data analysis for the four classes of adjectives tested.

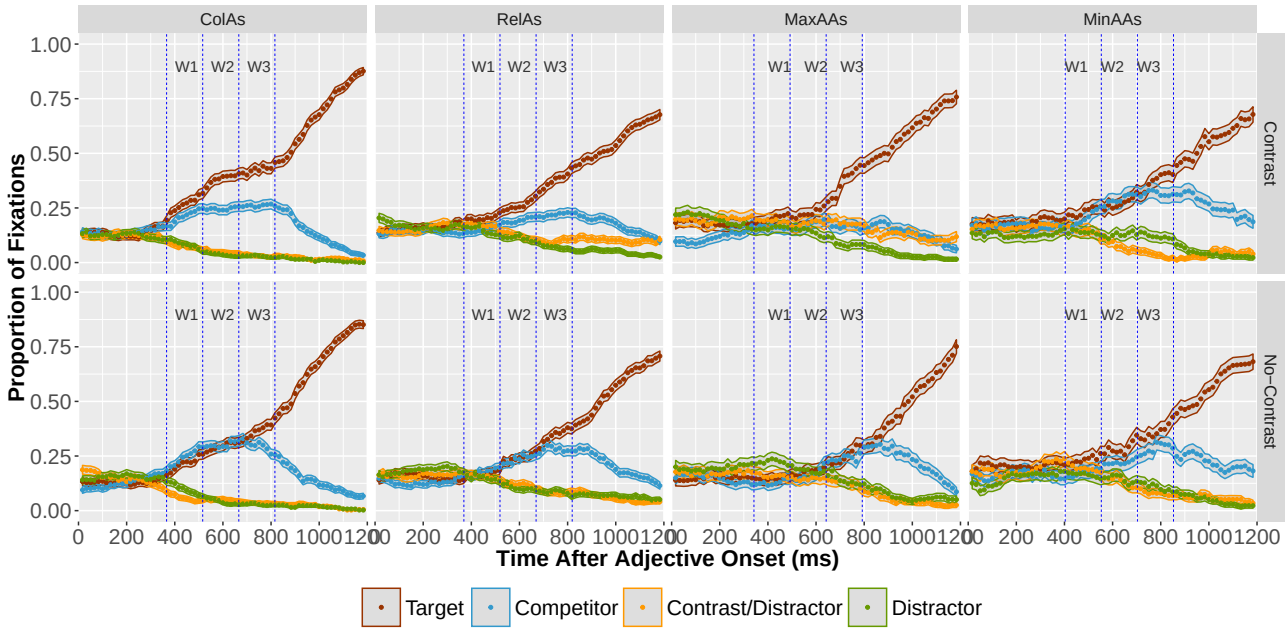


Figure 3.12: Proportions of fixations to each of the four objects in the display over time starting at the adjective onset (including the 200ms offset) for each adjective type. The vertical dashed blue lines mark the boundaries of the three windows defined for data analysis.

In order to determine whether target vs. competitor disambiguation occurred faster in the Contrast than in the No-Contrast condition, a one-way ANOVA using OBJECT TYPE (target vs. competitor) as a factor was run for each time window and adjective type. Given that the interest of this analysis lies in the timing differences between the two conditions tested and in particular whether target identification took place *faster* in the No-Contrast condition, I will only report those effects

that appeared in the earliest window for all adjective types. I first start with the results for **ColAs**. For this adjective type, the first significant effect in the Contrast condition takes place in W2 ($F(1,39) = 7.97$, $p < 0.008$), such that participants looked at the target object significantly more than they did to the Competitor. The effect does not reach significancy in the same time window of the No-Contrast condition ($F(1,39) = 0.12$, $p > 0.7$), suggesting that at this point participants had not yet discriminated between target and competitor. A two way ANOVA using OBJECT TYPE and CONDITION as factors did not show a main effect of BJECT TYPE ($F(1,39) = 2.50$, $p > 0.1$) or CONDITION ($F(1,39) = 1.76$, $p > 0.1$). However, there was a significant OBJECT TYPE x CONDITION ($F(1,39) = 6.86$, $p < 0.02$).

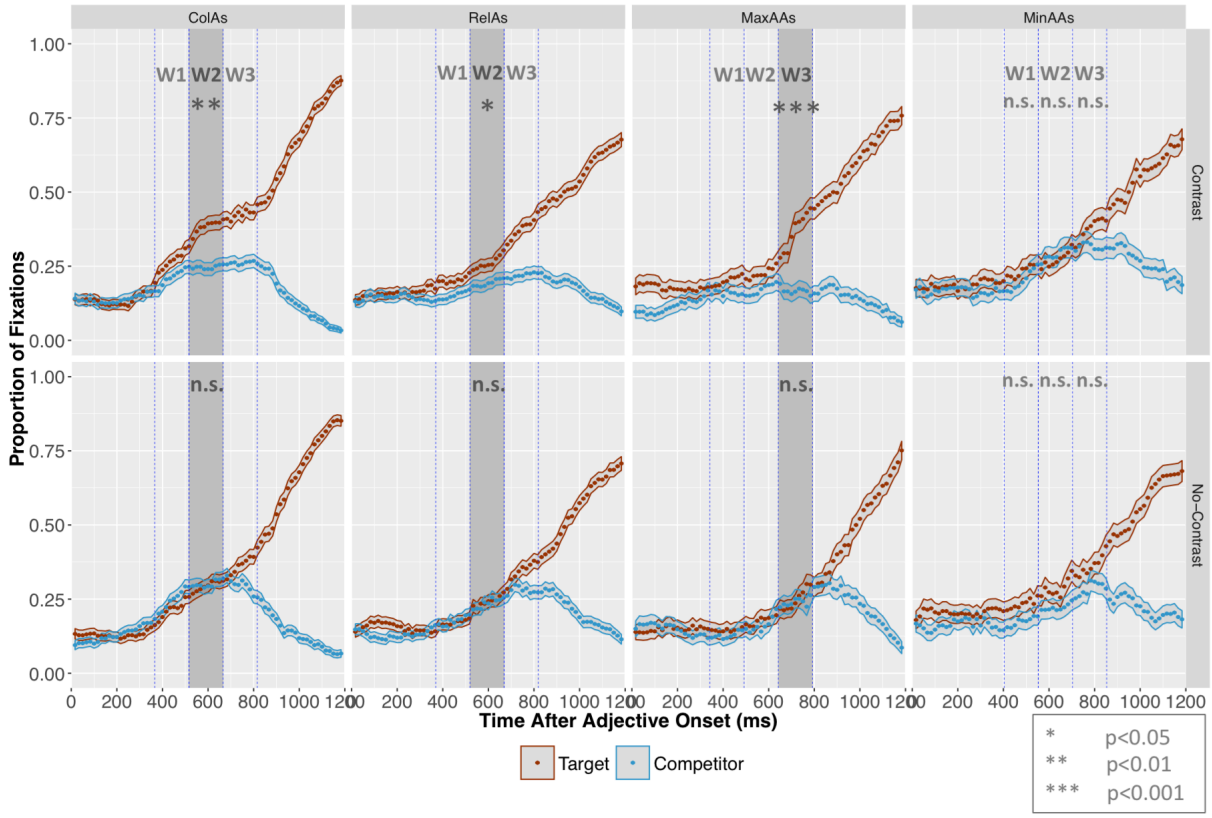


Figure 3.13: Proportions of fixations to target vs. competitor over time starting at the adjective onset. The grayed time windows correspond to the first time window in which a significant difference was found.

The same analysis procedure was followed with **RelAs**. As with ColAs, the results from a one-

way ANOVA run in each time window of the Contrast condition did not reach significance until W2 ($F(1,39) = 5.87$, $p < 0.05$). Again, the effect is driven by an increase in looks to the target object in this window. The analysis of W2 in the No-Contrast condition was not significant ($F(1,39) = 5.89$, $p > 0.9$), revealing that at this point participants had not yet committed to the target object. A two-way ANOVA did not show any significant main effects (OBJECT TYPE: $F(1,39) = 3.08$, $p > 0.8$; CONDITION: $F(1,39) = 0.55$, $p > 0.4$) or OBJECT TYPE x CONDITION interaction ($F(1,39) = 2.61$, $p > 0.1$).

A one-way ANOVA reveals that the first significant effect for **MaxAAs** in the Contrast condition does not appear until W3 ($F(1,39) = 13.82$, $p < 0.001$), whereas the effect in W3 of the No-Contrast condition is not significant ($F(1,39) = 0.03$, $p > 0.8$). As it was the case for ColAs and RelAs, this difference resulted from an increase in fixations to the target object in the Contrast condition. A two-way ANOVA showed a main effect of OBJECT TYPE ($F(1,39) = 9.14$, $p < 0.005$) that was driven uniquely by the changes in the Contrast condition, as confirmed by the results from the one-way ANOVA, and the fact that the OBJECT TYPE x CONDITION interaction was also significant ($F(1,39) = 4.28$, $p < 0.05$).

Regarding **MinAAs**, results did not reveal any significant main effect of CONDITION in any of the time windows examined (all p's > 0.1). W1 and W2 did not show a significant main effect of OBJECT TYPE (all p's > 0.1). The main effect of OBJECT TYPE reached significance in W3 ($F(1,39)=4.12$, $p < 0.05$). However, pair comparisons between target vs. competitor for the Contrast and No-Contrast conditions separately did not yield any significant results (p's > 0.1). No interactions between OBJECT TYPE and CONDITION (all p's > 0.3) were observed in any of the three windows. To verify whether there were any RECs in even later time windows, a fourth 150 ms window (W4) spanning from 856-1006 ms was examined. As in W3, a two-way ANOVA showed a main effect of OBJECT TYPE ($F(1,39) = 31.00$, $p < 0.00001$), but no significant main effect of CONDITION ($F(1,39) = 1.31$, $p > 0.2$), or OBJECT TYPE x CONDITION interaction ($F(1,39) = 0.47$, $p > 0.4$) was observed. A one-way ANOVA with OBJECT TYPE as factor revealed a significant differ-

ence between the two levels for both the Contrast ($F(1,39) = 12.59$, $p < 0.002$) and the No-Contras condition ($F(1,39) = 26.43$, $p < 0.00001$) such that participants fixated significantly more on the target object than the competitor object in both conditions.

Analysis II: Target vs. Target Comparison

In addition to the ANOVA analysis reported in the previous section, a second analysis using logistic mixed effects models was also performed. The goal of this analysis was to determine whether there were significant differences in the rate at which the proportions of fixations to the target objects in the Contrast and the No-Contrast conditions increased as a function of time. Therefore, the results of this analysis will help determine whether the effects reported in Analysis I were driven by an increase in looks to the target in the Contrast condition, or a decrease in looks to the competitor in the same condition.

Figure 3.14 plots the proportion of fixations over time to the target objects in the two conditions tested. The existence of a significant difference, such that the target object in the Contrast Condition received a higher proportion of looks earlier than the target object in the No-Contrast condition would be indicative of a REC. In order to identify any potential RECs, twelve time windows of 100 ms each starting at the onset of the adjective (offset by 200 ms) were defined for data analysis. The data from each of this twelve 100 ms windows was fit to a logistic mixed effect regression model to predict the proportions of fixation to the target object. The factor CONDITION was used as a main predictor. Random effects and intercepts for SUBJECTS and ITEMS were also specified in the model.

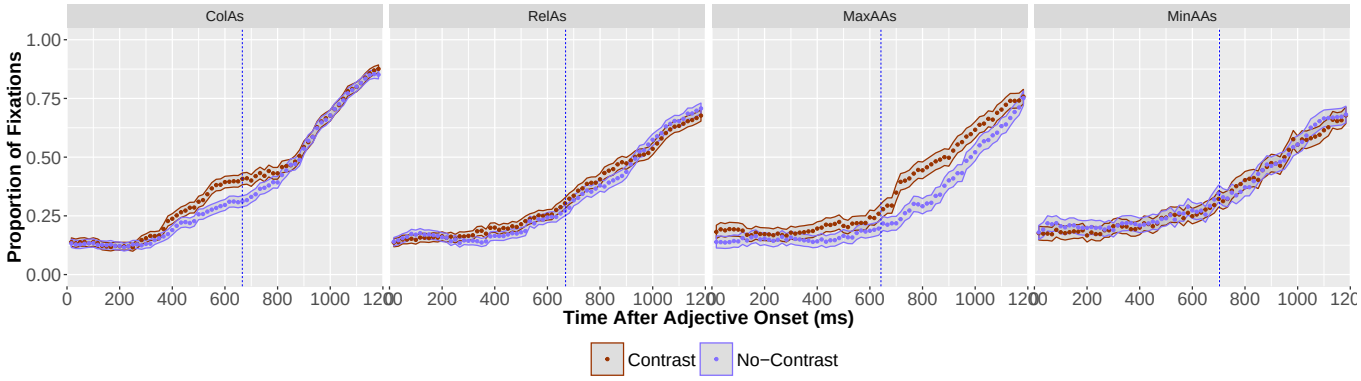


Figure 3.14: Proportions of fixations over time to the target objects in the Contrast and the No-Contrast condition. The plotted window starts at the adjective onset and spans for 1200 ms. The dashed lines mark the noun onset.

Figure 3.15 plots the z scores extracted from the models output for the main effect of CONDITION in each of the twelve time windows analyzed.[8] I use a criterion of $z > |2|$ to determine when there is a significant effect.

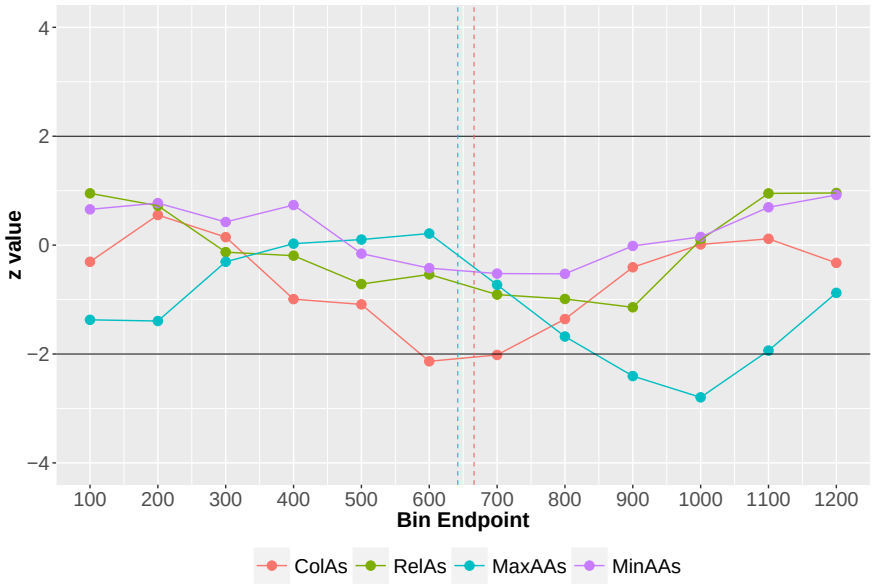


Figure 3.15: z scores from a logistic mixed effects model of eye movement data comparing looks to the target object in the Contrast and the No-Contrast condition for each of the twelve 100 ms time windows created for data analysis starting at the onset of the adjective. Each solid line represents one of the four adjective types tested. The dashed vertical lines indicate the onset of the head noun for the two classes of adjectives that displayed values > | 2 | in at least one time window, i.e. ColAs (red) and MaxAAs (blue).

The patterns of effects found in the target comparison analysis for ColAs, MaxAAs and MinAAs mostly mirror those found in the target vs. competitor disambiguation analysis reported in §3.2.3, although with some important timing differences. For **ColAs** a significant effect was found in the time windows spanning from 500 to 600 ms ($z = -2.134$), and from 600 to 700 ms ($z = -2.017$). These two time windows roughly correspond W2 (517-667 ms) in the analysis presented in §3.2.3, where a significant effect of target vs. competitor disambiguation was reported for ColAs.

In the case of **MaxAAs**, the target comparison analysis revealed effects in two consecutive time windows: the window spanning from 800 to 900 ms ($z = -2.405$), and the window spanning from

8. I follow Kingston *et al.* (2016) in reporting effects from Visual World eye-tracking data in the form of z-scores.

900 to1000 ms ($z = -2.796$). These effects are delayed with respect to the time window in which the first significant effect was found for the target vs. competitor disambiguation analysis, i.e. W3 (644-794 ms), which is the earliest window reflecting processing cost pertaining to the adjectival predicate. As can be observed in Figure 3.15, the window that stretches from 600 to 700 is precisely the window where the numerical increase in the absolute value of the z scores starts, although as stated before the effect does not reach significance until window 800 to 900 ms. Finally, results for **MinAAs** parallel those found in the target vs. competitor disambiguation results in that no significant effects were attested (see Figure 3.15).

Results pertaining to **RelAs** present the biggest disparity between the target vs. competitor disambiguation analysis and the target vs. target comparison analysis. While the former analysis showed a weak effect such that participants fixated significantly more on the target object than in the competitor object in W2, the latter analysis did not reveal any significant effects for this adjective class in any of the twelve time windows analyzed.

Analysis III: Adjectival Property Identification in the Absence of Contrast

The goal of this third analysis is to establish a baseline for the processing of the adjectival predicate in the absence of contextual support. Thus, this analysis focuses on the No-Contrast condition exclusively. Specifically, for each adjective type, I examine how quickly participants started to fixate significantly more on those objects that could be described by the adjectival property (i.e. the target and the competitor), as opposed to those objects that could not (i.e. the two distractors). The objective of this analysis is to explore the timing differences among adjective types between the time it took participants to isolate those objects bearing the adjectival property in the No-Contrast condition and the RECs described in §3.2.3. Therefore, the current analysis does not include MinAAs, as they did not display a REC.

For this analysis, I collapsed fixations to target and competitor on the one hand, and fixations to the two distractors on the other. Figure 3.16 plots the grand average of the proportions of fixations

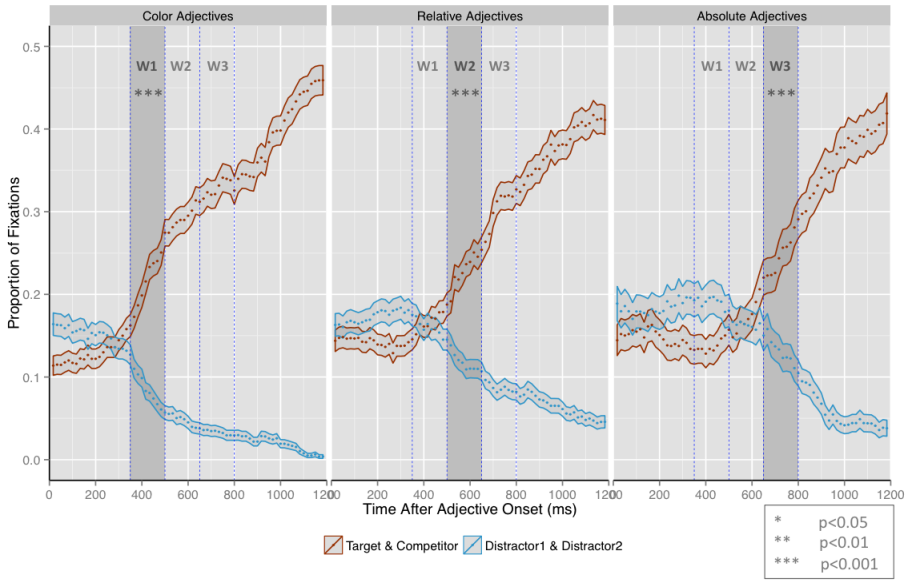


Figure 3.16: Proportions of fixations to target/competitor vs. distractors (No-Contrast condition) over time starting at the adjective onset. The grayed time windows correspond to the first time window in which a significant difference was found.

to target/competitor vs. two distractors.

As seen in Figure 3.16, visual identification of the adjective property (looks at target and competitor) is faster for ColAs than for RelAs and MaxAAs. A series of ANOVA's run on each time-window and each adjective type showed that in the case of ColAs, participants' looks to the target and competitor start to significantly diverge from looks to the two distractors in W1 ($F(1,39) = 55.57$, $p < 0.001$). The first significant effect for RelAs takes place in W2 ($F(1,39) = 43.50$, $p < 0.001$), whereas for MaxAAs, the first significant effect does not take place until W3 ($F(1,39) = 17.92$, $p < 0.001$). The effect remains significant in all subsequent windows for both ColAs (W2 and W3) and RelAs (W3).

3.2.4 *Interim summary of Effects*

Before proceeding to the discussion of the results found in Experiment 1, I summarize the significant effects for each of the three analyses carried out in the preceding sections (see Table 3.3). Analysis I consisted of a comparison of the data belonging to the target and competitor objects in both the Contrast and the No-Contrast conditions. In Analysis II, I compared the eye fixations to the target object in the Contrast and the No-Contrast condition. Significant effects resulting from analyses I and II can be taken to constitute evidence of RECs. Finally, analysis III zooms in into the No-Contrast condition in order to determine at what point participants fixated significantly more in those objects in the display that bore the adjectival property, i.e. target and competitor. The results of this analysis can be taken as a measure of how long it took for participants to process the lexical meaning of the adjective in the absence of contextual support for the restrictive interpretation.

	ColA	**RelA**	**MaxAA**	**MinAA**
Target vs. Competitor (Analysis I)	W2	W2	W3	n.s.
Target vs. Target (Analysis II)	500-600 ms	n.s.	800-1000 ms	n.s.
Prop. Ident. No-Cont. Cond. (Analysis III)	W1	W2	W3	——

Table 3.3: Table of time windows in which a significant effect emerged for the four adjective types and the three analyses performed on the data from Experiment 1. A green background indicates that the effect took place during the noun window, whereas a red background indicates that the effect took place in the noun window.

In the following section (§3.2.5), I discuss the relevance of these findings in light of the research questions that the experiment sought to address.

3.2.5 *Discussion*

Results from Analysis I showed that ColAs, RelAs and MaxAAs gave rise to RECs, since target identification occurred faster in the Contrast Condition than in the No-Contrast condition, whereas no REC was detected for MinAAs. Importantly, the timing and the strength of the attested RECs varied across adjective type. In this section, I will start by describing these differences, as well as

discussing the implications that they have for the theories of the relative vs. absolute distinction introduced in Chapter 2. Next, I will discuss the results from Analysis III and will argue that at least two different RECs should be distinguished based on the role that the competitor object has in the Contrast condition. I will furthermore argue that this difference might be taken to be the processing blueprint that allows to tease apart RECs that are uniquely driven by pragmatic reasoning about informativity from RECs that involve other types of context-sensitivity associated with the grammatical and/or pragmatic properties of a given predicate. Finally, I will conclude with a discussion about the puzzling lack of context-sensitivity displayed by MinAAs.

Results from Analysis I show that not all RECs took place during the same time window: ColAs and RelAs displayed RECs during the adjective window (W2), whereas MaxAAs did not give rise to a REC until the beginning of the noun window (W3). This is a significant qualitative difference that sets MaxAAs apart from RelAs and ColAs. In the latter case, information of the visual context was incorporated early and used predictively to make an educated guess about the identity of the target, even before linguistic information about the head known was available to the participants. However, in the case of MaxAAs the same contextual information was integrated late, such that participants did not make decisions about the identity of the target object until they had sufficient linguistic information, i.e. linguistic input about the head noun, to do so. Even though ColAs, RelAs and MaxAAs all showed sensitivity to the contrastive visual information, the processing of MaxAAs displayed a dependency of the head noun that was not paralleled by RelAs and ColAs.

The second axis of variability in the RECs uncovered by Analysis I was in the strength of the effects. Results from this analysis revealed effects of bigger magnitude for ColAs and MaxAAs than for RelAs, which also failed to give rise to an OBJECT TYPE x CONDITION interaction. The weakness of the effect found for RelAs is unexpected given the results from Sedivy *et al.* (1999), who report a robust REC for this same class of adjectives. The cause for this difference may lie in the different properties of the competitor objects used in both studies. As noted in §2.3.1, the competitor objects used by Sedivy *et al.*'s had a higher degree of the adjectival property than

the target, but could not be described by the positive form of the adjective. For instance, if the target object was a tall glass, the competitor object could be a *taller* picture that was not itself tall. In contrast, in the current study the competitor objects always matched the target with respect to the degree to which they possessed the adjective property. Therefore, it is possible that the competitors used in this experiment attracted a higher proportion of fixations in the early stages of processing of the adjectival predicate, thus weakening the REC. Indeed, a close examination of Sedivy *et al.*'s results confirms that participants stopped fixating on the competitor object as early as 200 ms after the adjective onset (see Figure 2.7 in Chapter 2). Therefore, it is likely that in the current experiment participants coerced a Comparison Class during the processing of the adjective window that included the contrast set *and* the competitor object, leading to the observed increase in the proportion of fixations to the competitor.[9] In fact, Leffel *et al.*, who made use of the same visual stimuli tested in the current study, argue that their results show evidence of such coercion. As was discussed in §2.3.1, Leffel *et al.* did not detect a REC for RelAs when the target object had a lower degree of the adjectival property than the competitor, suggesting that participants treated the competitor as a highly likely referent of the adjectival phrase during the early stages of processing, despite the presence of a contrast set in the visual display. In contrast, in Experiment 2, where the target object was a better exemplar of the adjectival property than the competitor, target identification in the Contrast condition took place early in the adjective window. However, the REC was not detected until later, well into the noun window, apparently due to the fact that the No-Contrast condition also registered a higher proportion of looks to the target object in the adjective window. This increase in fixations to the target object in the No-Contrast condition during the adjective window is not unexpected given that the target object instantiated the adjectival property to a higher degree than the competitor and was therefore considered a more probable target

9. Alternatively, it could also be the case that instead coercing one single Comparison Class, participants maintained to parallel competing Comparison Classes: one containing the target object plus the contrast object (in the Contrast condition), and one consisting of a singleton set containing the competitor object. This would also predict the higher proportion of looks to the competitor object observed in the data.

candidate than the competitor object.

Given the preeminent role that competitor objects seem to have in the current results, an important question is whether the RECs found in Analysis I were driven by an increase of looks to the target object in the Contrast condition, or a decrease of looks to the competitor object in the same condition. This question can be partially addressed by comparing the results from Analysis I and Analysis II. In particular, if the same time window in which Analysis I revealed a REC shows no significant effect in the target vs. target comparison analysis, this would indicate that the REC must have mostly resulted from a decrease in fixations to the competitor object in the Contrast condition. For ColAs, the REC took place in W2 (517-667 ms). The only significant effect for ColAs in Analysis II took place precisely in the time window spanning from 500-600 ms, thus suggesting that in the case of ColAs, the REC resulted at least partially from an increase in looks to the target object. In the case of MaxAAs, the REC was detected in W3 (644-794 ms). However, the target vs. target comparison did not show significant effects during the 600-800 time windows. Therefore, I conclude that for MaxAAs the competitor, not the target object object, was the main driver of the REC detected in W3. This was confirmed by a post hoc analysis where looks to the competitor objects belonging to the Contrast and the No-Contrast condition were compared using a logistic mixed effects regression model parallel to the one used in Analysis II with the Contrast condition coded as the reference level. Even though the window spanning from 600 to 700 ms did not show a significant difference between the two competitor objects ($\beta = 0.2857$, $SE = 0.520$, $p > 0.5$), the subsequent window spanning from 700 to 800 ms did show a significant effect such that participants fixated more on the competitor object in the No-Contrast condition compared to the Contrast condition ($\beta = 1.0513$, $SE = 0.521$, $z = 2.015$, $p < 0.04$). Regarding RelAs, Analysis II did not reveal any significant effects in any of the twelve analyzed time windows, suggesting that the weak REC detected in W2 (520-670 ms) by Analysis I should have resulted from a decrease in looks to the competitor object in the Contrast condition rather than an increase in looks to the target in the same condition. A post-hoc analysis in the time window spanning from 500-600 ms

and 600-700 ms, comparing fixations to the two competitor objects did not reveal any significant effects (all p's > 0.1). I also run the same model on the data corresponding to W2, as defined for Analysis I. Results from this last analysis showed a statistical trend for this comparison ($\beta = 0.4105$, $SE = 0.2432$, $z = 1.7$, $p > 0.9$). Taken together, results from Analyses I and II suggest that the form of the RECs displayed by RelAs and MaxAAs was very similar, albeit results displayed by RelAs were less robust than those observed for MaxAAs: for these two classes of adjectives RECs came about as a result of a faster target vs. competitor disambiguation in the Contrast condition vs. the No-Contrast condition that was mainly driven by a decrease in looks to the competitor object in the Contrast condition. The REC displayed by ColAs showed a different processing blueprint: the effect also resulted from a faster disambiguation between the target and the competitor in the Contrast condition compared to the No-Contrast condition. However, in this case the target object had a significant role in driving the effect.

I now turn to the results from Analysis III. Results from the No-Contrast condition alone provide us with a baseline of how quickly participants can identify different types of adjectival properties (in a visual display) in the absence of contextual support for the restrictive interpretation. The results show that the color properties were identified faster (W1) than properties introduced by RelAs (W2) and MaxAAs (W3). By comparing the timing of the effects found in Analysis III to the RECs found in Analysis I, it is possible to determine whether the speed with which the target object was identified in the Contrast condition was the same relative to the time it took for participants to identify the adjectival property in the No-Contrast condition. The goal of this comparison is to obtain a finer-grained picture of the timing properties of the different RECs not by comparing them across adjective types (as was done in Analysis I), but rather by quantifying the processing advantage found in the Contrast condition over the No-Contrast condition within each adjective type. In comparing results from Analysis I and III, there exist at least two possible scenarios based on the visual inspection of the current results: first, the effects from each analysis could align, i.e. they might occur in the same time window; second, the REC might take place

after the adjectival property was disambiguated in the No-Contrast condition. The former pattern would entail a greater processing advantage in the Contrast condition compared to the No-Contrast condition, because as soon as participants were able to first zoom into the adjectival property in the No-Contrast condition, they had already disambiguated between the target and the competitor object in the Contrast condition. The latter pattern would indicate a smaller processing advantage of the Contrast condition over the No-Contrast condition, since the REC would be delayed with respect to the time in which participants were able to to first zoom into the adjectival property in the No-Contrast display. The first pattern of results is exactly what was found for RelAs and MaxAAs, for which all effects take place in the same time window (W2 for RelAs and W3 for MaxAAs). ColAs, on the other hand, displayed the second pattern of effects: while the adjective property had already been identified in the No-Contrast condition by W1, the REC was not significant until W2.

As discussed above, the only REC that showed an increase in fixations to the target object was the REC displayed by ColAs, while RECs triggered by RelAs and MaxAA were uniquely driven by a significantly lower proportion of fixations in the Contrast condition compared to the No-Contrast condition. Given the timing differences uncovered by Analysis III, an important question becomes whether the competitor object was considered as a distinctive potential referent on par with the target object at any time before the emergence of the REC in the Contrast condition. In order to address this question, I conducted a post-hoc analysis where I analyzed the window immediately preceding the REC for each adjective type, i.e. W1 for ColAs and RelAs and W2 for MaxAAs. The objective was to determine whether in the window that immediately preceded the REC in the Contrast-condition, participants had already distinguished those objects that bore the adjectival property (i.e. target and competitor) from those that did not (i.e. contrast and distractor), or whether they were still looking equiprobably at these two sets of objects. As in Analysis III, I collapsed looks to the target and competitor on the one hand, and looks to the contrast object and the distractor on the other, and I fit a logistic mixed effects regression model parallel to the one used in Analysis III to the data belonging to the relevant time windows and adjective type. Findings again show that

ColAs patterned differently from RelAs and MaxAAs: a main effect of OBJECT TYPE was found for ColAs such that participants fixated significantly more on the target and competitor objects than to the remaining two objects in the display by W1 ($\beta = 1.5327$, $SE = 0.2587$, $z = 5.925$, $p < 0.001$). No effect was detected for RelAs in W1 ($\beta = 0.2384$, $SE = 0.2524$, $z = 0.945$, $p > 0.3$), or for MaxAAs in W2 ($\beta = 0.4372$, $SE = 0.4904$, $z = 0.891$, $p > 0.3$). These findings suggest that in the early stages of the adjective window, the competitor object was taken to be a plausible referent of the modified Noun Phrase despite the presence of the contrast object only in the case of ColAs, whereas in the case of RelAs and MaxAAs the presence of the contrast set suppressed participant's attention to the competitor. These findings are consistent with the conclusions extracted from the comparison of Analyses I and II, which suggested that the competitor object was discarded early during the processing of RelAs and MaxAAs, since for these adjectives the REC resulted from lower proportions of looks to the competitor object in the Contrast condition compared to the No-Contrast condition.

Two types of RECs

One of the main goals of Experiment 1 was to tease apart RECs that only involve Gricean pragmatic reasoning from those that recruit other types of pragmatic and/or lexical context-sensitivity. The current results, summarized in Table 3.3 (repeated below as Table 3.4) and Table 3.5, point at two different types of RECs based on whether: a) the target object was the main driver of the REC (Analyses 1 and 2 in Table 3.5), and 2) whether participants zoomed into the adjectival property in both the Contrast and the No-Contrast condition in the window immediately preceding the REC (Analyses 3 and 4 in Table 3.5).

Given the two properties defined above, Table 3.5 shows that ColAs patterned differently from RelAs and MaxAAs: for ColAs, the REC was mostly driven by a higher proportion of looks to the target object in the Contrast condition compared to the No-Contrast condition. Furthermore, the REC was more progressive in the sense that immediately before target vs. competitor disambigua-

	ColA	RelA	MaxAA	MinAA
Target vs. Competitor (Analysis I)	W2	W2	W3	n.s.
Target vs. Target (Analysis II)	500-600 ms	n.s.	800-1000 ms	n.s.
Prop. Ident. No-Cont. Cond. (Analysis III)	W1	W2	W3	——

Table 3.4: [Table 3.3 repeated] Table of time windows in which a significant effect emerged for the four adjective types and the three analyses performed on the data from Experiment 1. A green background indicates that the effect took place during the noun window, whereas a red background indicates that the effect took place in the noun window.

	ColAs	RelAs	MaxAAs
1. Effects Analyses I & II occurred in the same window	yes	——	no
2. Competitor vs. Competitor comparison significant in window where REC appeared	——	yes (trend)	yes
3. Effects Analyses I & III occurred in the same window	no	yes	yes
4. Target/Competitor vs. Contrast/Distractor disambiguation immediately before REC	yes	n.s.	n.s.

Table 3.5: Summary of results obtained from comparing Analyses I & II (1) and Analyses II & III (3), as well the results from the different post hoc analyses (2 & 4) performed on the data from Experiment 1.

tion took place in W2 of the Contrast condition, participants had considered both the target *and* the competitor objects as potential referents of the modified noun phrase. RelAs and MaxAAs behaved differently, since they gave rise to RECs that mainly resulted from smaller proportions of looks to the competitor object in the Contrast condition compared to the No-Contrast condition. Moreover, in the Contrast condition participants never considered the competitor object as a plausible referent along with the target object after discarding the other two objects in the display. Based on the current results, I would like to argue that these two features are a good diagnostic to distinguish RECs that are fueled by global pragmatic reasoning related to informativity (ColAs) from RECs that also

involve other classes of context-sensitivity (RelAs and MaxAAs).[10]

MinAAs

MinAAs were the only adjective type that did not present a REC. Importantly, the lack of a REC for MinAAs cannot be attributed to the inability on the part of the participants to identify the objects bearing the adjectival property in the visual display. Figure 3.17 plots the proportions of fixations in the Contrast condition to the two objects that could be described by the adjectival property (i.e. target and competitor), and the two objects that could not (i.e. contrast and distractor). As can be seen in figure 3.17, participants discriminated between target/competitor vs. contrast/distractor as early as W1 of the adjective window. This was confirmed by a one-way ANOVA run on W1 and W2 of the Contrast condition after collapsing fixations to the target and the competitor objects on the one hand, and looks to the contrast and distractor on the other. Results from W1 showed a significant effect of OBJECT TYPE such that the proportion of looks to the target & competitor objects where significantly higher than the proportion of looks to the other two objects in the display ($F(1,39) = 7.00$, $p < 0.02$). In W2, the effect did not only remained significant but in fact became stronger ($F(1,39) = 46.467$, $p < 0.0001$).

The early identification of the adjectival property in the Contrast condition suggests that participants indeed recognized and processed the meaning of the adjective very early during the adjective window, unlike what was observed for MaxAAs, which nevertheless gave rise to a REC. Despite the early identification of the MinAA property in the Contrast condition, target vs. competitor disambiguation did not take place until W4. More importantly, early identification of the adjectival property in the Contrast condition did not lead to a faster identification of the target object in the Contrast condition compared to the No-Contrast condition, reinforcing the conclusion that contextual information about contrast is not relevant for the interpretation of MinAAs. In the following

10. In §3.2.5, I will argue that within the latter class of RECs, different types of context-sensitivity can further shape the timing properties of the REC.

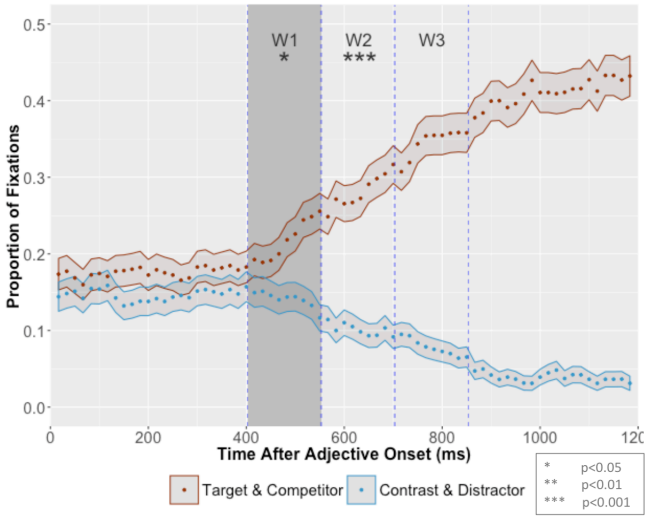


Figure 3.17: Proportions of fixations to Target & Competitor vs. Contrast & Distractor over time starting at the adjective onset for the Contrast condition for MinAAs. The grayed time window corresponds to the first time window in which a significant difference was found.

sections, I discuss the significance of these findings in light of the predictions made by the different theories if the relative/absolute distinction presented in Chapter 2.

The relative vs. absolute distinction

The second goal of Experiment 1 was to address the question of whether contextual information is recruited in comparable ways during the processing of RelAs and AAs. In §3.2.1, I have argued that theories that treat all gradable thresholds as context-sensitive, such as Lassiter & Goodman's proposal where the relative/absolute distinction is *de facto* blurred, predict that, all things being equal, information from the visual context should be integrated in comparable ways for all GAs. However, the current results speak against this homogeneous view of GAs. Despite the fact that the processing of both RelAs and MaxAAs showed sensitivity to context, participants incorporated information about context at very different stages of processing. In the case of RelAs, our results replicate Sedivy et al.'s (1999) findings in that participants were able to integrate information from the visual context early during the adjective window. This suggests that during the processing of

RelAs, contextual information about contrast served to efficiently predict the identity of the target object at a point where the linguistic input was still compatible with two objects in the display (i.e. the target and the competitor). For MaxAAs, contextual information also facilitated target identification in the Contrast condition, but this information was not integrated until the noun window, suggesting that in this case participants heavily relied on the head-noun information to resolve the referent of the definite NP. Perhaps more problematic for Lassiter & Goodman's account are the results pertaining to MinAAs. The lack of sensitivity to context displayed by this class of adjectives is more difficult to reconcile with a theory that argues for context-sensitive gradable thresholds across the board, specially given the fact that MaxAAs, which in this account form a minimal pair with MinAAs, did show context-sensitivity effects.

Taken together, the current results suggest that there exists a three-way distinction in the context-sensitivity patterns displayed by GAs. How to derive these differences from Lassiter & Goodman's RSA model is however not obvious, since in this system the distinctions among GAs are not grammatically encoded in the first place, and the only source of variability across GAs that could be responsible for the observed differences is in the specifications of the interpreter's background knowledge.

The distinctions among GAs detected in Experiment 1 can be better explained by theories that assume shifting relative thresholds and fixed absolute thresholds. As discussed in chapter 2, this brand of theories ground the context-sensitivity of RelAs to the lexical representation of the predicate, while the context-sensitivity of MaxAAs is attributed to pragmatic reasoning about imprecision. I would like to argue that the lack of a REC for MaxAAs in the adjective window coupled with the fact that target identification did not take place until the noun window is better accounted for by the view that assumes context-insensitive thresholds for MaxAAs. The fact that participants did not make use of the contextual information about contrast in the adjective window of MaxAAs indicates that the linguistic auditory instruction was perceived as truly ambiguous. This contrasts with the results observed for RelAs, for which the presence of the Comparison Class in the visual

display was sufficient to overcome this ambiguity.

An independent question is how or whether imprecision calculation should account for the target identification facilitation observed in the noun window of the MaxAAs condition. The current experimental design does not explicitly address this question, but I would like to make a couple of remarks regarding the potential role of imprecision calculation in the present results. The first remark is that even though the visual stimuli used in Experiment 1 were compatible with a precise interpretation of the MaxAAs, it would be premature to conclude that participants did not perform any sort of imprecision calculation during the processing of the modified NPs. One possibility is that imprecision calculation is only performed whenever the truth-conditions need to be relaxed in the context (Kennedy 2007; Leffel *et al.* 2016). However, it has also been argued that the precision threshold is a pragmatic parameter that must always be fixed in the interpretation of a MaxAA, even when the precise interpretation is available (van Rooij 2011). Therefore, imprecision calculation cannot be ruled out as one of the possible drivers of the REC displayed by MaxAAs in the current results. The second remark is that despite being compatible with a precise interpretation, the visual stimuli used in the experiment were also compatible with lower precision thresholds, even the lowest one. Therefore, the current experiment does not rule out the possibility that participants reasoned about imprecision and/or adopted a low precision thresholds for the interpretation of MaxAAs trials.[11] Perhaps a more pressing question is what precise mechanisms would be responsible for facilitating imprecision calculation in the Contrast condition compared to the No-Contrast condition. One option is that the contrast object in the Contrast condition offered a richer range of degrees in the relevant adjectival scale that facilitated the computation of the precision threshold. Further research must be conducted in order to determine whether this speculation is on the right track.

11. In Chapter 4, I discuss in more detail approaches to imprecision calculation in the context of numerals that in fact argue that there exists a bias towards imprecision such that imprecise interpretations are defaulted to whenever possible (see §4.4).

Finally, The lack of a REC for MinAAs cab be easily accommodated by theories that assume fixed absolute thresholds, if, as I have argued, MinAAs do not involve imprecision calculation and its interpretation is not context-dependent. The fact that MinAAs did not pattern with ColAs in the current results also suggests that not all prenominal adjectives trigger reasoning about contrast, as would be expected by the naive pragmatic view.

A potential confound

The previous discussion does not address the possibility that different patterns of interaction between RelAs and MaxAAs could be due to differences in the expectations about contrastive use with which each of these adjective types is associated. As the results pertaining to MinAAs show, not all adjective classes trigger reasoning about contrast. This finding argues against a naive pragmatic view of informativity, and points at the existence of a potential confound, namely that the differences between ColAs, RelAs and MaxAAs observed in Experiment 1 could be fueled by differences in their overspecification penalty. For instance, it is conceivable that bigger overspecification penalties might give rise to earlier RECs. This is an important point when interpreting the current results, especially with respect to the differences between RelAs and MaxAAs, for which we are mostly interested in any potential differences in context-sensitivity beyond general quantity and manner-based pragmatic reasoning. In order to address this concern, Experiment 2, presented in the following section, quantifiers the overspecification penalty of each of the adjective classes tested in Experiment 1, with the goal of determining whether informativity alone can account for the results obtained in Experiment 1.

3.3 Experiment 2: Perceived Informativity

3.3.1 Perceived Informativity and Referential Effects of Contrast

Despite the fact that RECs have been consistently replicated with adjectivally modified NPs (Sedivy *et al.* 1999; Sedivy 2004; Weber *et al.* 2006; Grodner & Sedivy 2011; Wolter *et al.* 2011; Aparicio *et al.* 2015; Leffel *et al.*), the exact mechanisms underlying these effects are not fully understood, and it remains an open question whether all the RECs reported in the literature are born equal (cf. Sedivy 2003, 2004) implicate the same amount if at all of pragmatic reasoning or whether there are other sources. As in the current study, experiments that have demonstrated the existence of RECs minimally compare two conditions: The Contrast condition, which contains a contrasting set of objects that only differ with respect to the degree to which they posses the adjectival property such that only one of the two objects could be felicitously described by the adjective, and the No-Contrast condition, which lacks such contrasting set. Therefore, the crucial difference between the Contrast condition and the No-Contrast condition is that in the former, the visual display includes objects that contrast only with respect to the information provided by a noun modifier, not with respect to the information provided by the head noun; while in the latter all objects in the display contrast with respect to the information provided by the noun. This makes the use of a modifier non-contrastive or "redundant," since the head noun alone suffices to distinguish the intended referent from the other objects in the display. As discussed above, a REC is observed when visual target identification takes place significantly faster in the Contrast condition compared to the No-Contrast condition. Such effects receive a natural pragmatic explanation in terms of the interaction of the Gricean Maxims of Quantity and Manner (Grice 1975). Since a definite description with a restrictive modifier is both more complex and more informative than a corresponding description without a modifier, a speaker's use of a modified form provides an indication that she intends to refer to an object that contrasts relative to the modifier but not the noun, which in turn facilitates referential fixation in the Contrast condition but not in the No-Contrast condition.

A naive version of the Gricean account of RECs would lead to the expectation that (cooperative) uses of modifiers should be restricted to contexts involving contrast; i.e., contexts in which the modifier is not redundant, in the sense described above. However, there is evidence that speakers frequently use modifiers in referential NPs even in the absence of contrast (Pechmann 1989; Nadig & Sedivy 2002; Sedivy 2003; Maes *et al.* 2004; Sedivy 2004; Koolen *et al.* 2011). Certain patterns seem to emerge in the use of such apparently redundant adjectives. Experimental production tasks have consistently shown that color adjectives are more likely to be used redundantly than other classes of adjectives like dimensional or material adjectives (Pechmann 1989; Belke & Meyer 2002; Nadig & Sedivy 2002; Sedivy 2004). Several factors have been found to be good predictors of when a speaker is more likely to use a redundant adjective. For instance, color adjectives that denote a stereotypical property of the object (e.g., a *yellow* banana) are less likely to be used redundantly (Sedivy 2003), while atypical color adjectives are more likely to be used redundantly (Westerbeek *et al.* 2015). A second factor affecting the production of redundant adjectives in referential communication tasks is the amount of variation present in the visual scene. Speakers are more likely to utter an overspecified description when the visual scene contains color variability, i.e. the visual display is polychrome, than when it does not, i.e., the visual display is monochrome (Koolen *et al.* 2013; Rubio-Fernández 2016). Based on the results from the Visual World experiment pertaining to MinAAs we already know that a naïve version of the Gricean account cannot be on the right track, since this adjective class did not trigger a REC when used prenominally as a restrictive modifier.

The fact that speakers not only often choose to include overspecified adjectives as part of their utterances, but also do so in systematic ways is unexpected in the context of the naive Gricean view, in which all redundant adjectives are suboptimal from an informativity point of view. Rubio-Fernández (2016) suggests that overspecification should be recast in terms of efficiency rather than informativity, as modifiers may facilitate target identification by helping the hearer optimize the visual search of the target object (see Paraboni *et al.* 2007; Arts *et al.* 2011 for similar claims). In

this respect, efficiency can be regarded as a pragmatic cooperative phenomenon. Assuming that hearers are sensitive to the systematicities in the production patterns of redundant adjectives, different adjective classes could in principle be associated with different expectations regarding the probability that a given adjective will be used contrastively. This is relevant for VW experiments such as the ones discussed above, as it leads to a more nuanced prediction than the naive Gricean view, namely that only those adjective classes for which a redundant adjective is *perceived* as providing too much information in the context should give rise to such effects, i.e. there should be a correlation between perceived overinformativity and strength of referential contrast. The resulting picture, like the naive Gricean one, remains rooted in reasoning about (over-)informativity of a complex form, but allows for variation in classes of modifiers based on the extent to which they are independently perceived as over-informative or not.

Going back to Experiment 1, to test this hypothesis I conducted Experiment 2, where I compare all four classes of adjectives tested in the eye-tracking experiment presented above for perceived informativity. I show that minimum standard adjectival modifiers differ from all the other classes of adjectival modifiers in not being perceived as overinformative in the absence of contextual support for contrastive interpretations, a finding that provides evidence for the perceived informativity-based view of RECs described above. However, among the other three classes of adjectives, it is also found that the magnitude of the perceived (over)informativeness does not completely map to the size of the RECs reported in Experiment 1. I conclude with discussion of the role that lexical semantic factors may play in driving perceived informativity and variable RECs.

The same lists and adjectives (RelAs=9, MaxAAs=4, MinAAs=6, ColAs=4) used in the eye-tracking studies were tested (See Table 3.1), with a total of 60 experimental items (20 containing RelAs, 10 containing MaxAAs, 10 containing MinAAs and 20 containing ColAs). Conditions were distributed in two lists using a Latin Square design. Both the order of the trials within each list and the position of the four pictures within each trial were randomized (see Figure 3.18).[12] The same

12. The full list of the experimental items used in Experiment 2 can be found in the following link:

60 filler trials used in Experiment 1 were included (see Figure 3.2).

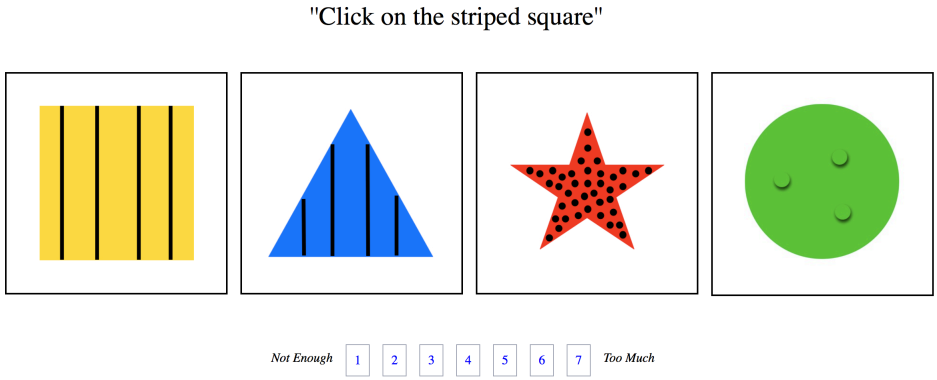


Figure 3.18: Item example for Experiment 2.

3.3.2 Predictions

Experiment 2 addresses the question of whether all the adjective types tested in Experiment 1 (i.e. ColAs, RelAs and AAs) are perceived as equally informative when the display contains a contrastive object (Contrast condition), compared to displays that do not (No-Contrast condition). With this goal in mind, Experiment 2 consists of an offline judgement task, where participants were requested to rate whether the instructions used in the eye-tracking experiment provide a sufficient amount of information to confidently identify the target object in the relevant visual display.

If the online eye-tracking effects reported above are shaped by differences in the perceived informativity, the following patterns of results for Experiment 2 are predicted: First, since MinAAs were the only type of adjective that did not give rise to REC, no differences in perceived informativity between the Contrast and the No-Contrast conditions should arise for this adjective class. All other adjectives should show a significant difference between these two conditions such that the No-Contrast condition is perceived as overinformative compared to the Contrast condition. Second, based on the timing of the RECs observed for ColAs, RelAs and MaxAAs, it is expected that the magnitude of the overspecification penalty should be greater for MaxAAs than for ColAs and

http://lucian.uchicago.edu/blogs/lpl/exp2_aparicio_dissertation/

RelAs.

3.3.3 Methods

Materials

Stimuli consisted of the same visual displays used in Experiment 1 (a total of 120). The auditory instructions used in the eye-tracking experiment were transcribed and accompanied the visual displays.

Procedure

Participants saw displays of four pictures on a computer screen coupled with a written statement such as *Click on the striped square*. For each of the displays, participants were instructed to rate whether the instruction provided a sufficient amount of information to identify the right target. Judgments were indicated on a 1-7 scale, where 1 corresponded to *Not enough information* and 7 corresponded to *Too much information*. At the beginning of the experiment, participants had three practice trials to help them become familiar with the task.

Participants

Participants were 32 native speakers of English between the ages of 18-35 (12 females; mean age = 30) recruited through the website Amazon Mechanical Turk. Three subjects were removed from data analysis because they were not between 18-35 leaving a total of 29 (10 females; mean age = 29). All participants were payed $3 and were naïve to the purpose of the experiment.

3.3.4 Results

Means were obtained for all adjective types. Visual inspection of the left plot in Figure 3.19 reveals that the No-Contrast condition received higher ratings compared to the Contrast condition

for ColAs, RelAs and AAs. For the class of AAs, data from MaxAAs and MinAAs were combined. The ratings in the Contrast condition were used as the baseline comparison against the ratings in the No-Contrast condition, as the former represents ratings pertaining to the condition containing the optimal amount of information, since target identification would not be possible in the absence of the adjective. Paired t-tests confirm that the differences between the two conditions were statistically significant (ColAs: $t(28) = -5.78$, $p < 0.0001$; RelAs: $t(28) = -3.20$, $p < 0.01$; AAs: adjectives $t(28) = -3.85$, $p < 0.001$). However, closer inspection to the two subclasses of AAs (central plot, Figure 3.19) shows that the difference between conditions observed for AAs is mostly driven by MaxAAs, which present the higher ratings in the No-Contrast condition. A paired t-test confirmed that this difference was highly significant ($t(28) = -5.89$, $p < 0.0001$). MinAAs, on the other hand, showed a non-significant difference across conditions ($t(28) = -0.91$, $p > 0.3$).

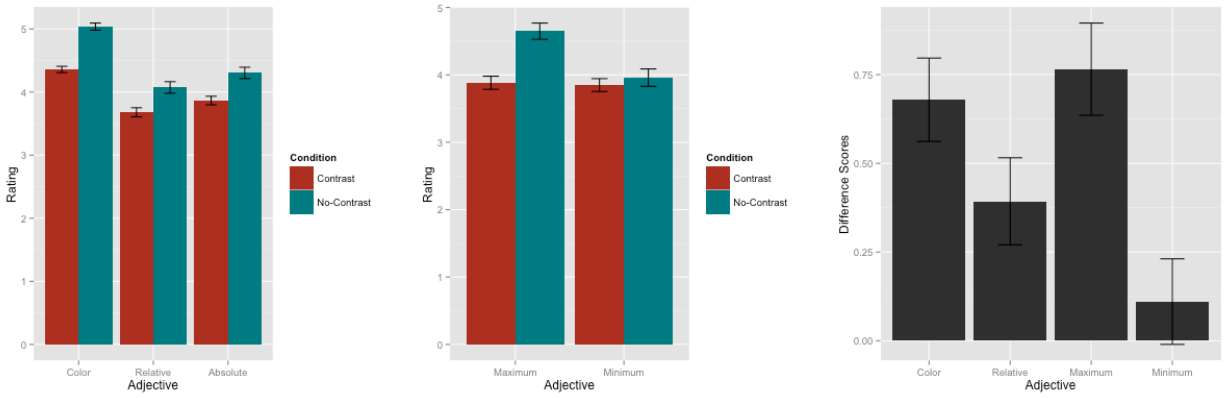


Figure 3.19: **Left.** Rating means for ColAs, RelAs and AAs; **Central.** Rating means for MaxAAs and MinAAs; **Right.** Difference Scores between the Contrast and the No-Contrast condition for each adjective type.

The data from the three adjective classes was submitted to a 2-way ANOVA using ADJECTIVE TYPE and CONDITION as factors, which showed significant differences between the two conditions, i.e. ColAs, RelAs and MaxAAs. A significant interaction for ADJECTIVE TYPE x CONDITION was detected ($F(2,56) = 7.64$, $p < 0.008$), showing that the magnitude of the effect was different across the three adjective types. In order to further explore this interaction, a 2-way ANOVA was run in three different subsets of the data. The interaction remained significant for the subset containing

RelAs and MaxAAs ($F(1,28) = 10.70$, $p < 0.002$), and the subset containing RelAs and ColAs ($F(1,28) = 13.10$, $p < 0.001$), while it did not reach significance for the data subset containing only ColAs and MaxAAs ($F(1,28) = 0.7$, $p > 0.4$). This suggests that the magnitude of the effect was comparable for ColAs and MaxAAs (see right panel of Figure 3.19 containing the difference scores obtained by subtracting the Contrast condition from the No-Contrast condition for each adjective type), and that the ADJECTIVE TYPE x CONDITION interaction detected for the full data set was driven by differences between ColAs and MaxAAs on the one hand and RelAs on the other.

3.3.5 Discussion

For ColAs, RelAs and MaxAAs, the No-Contrast condition received significantly higher ratings than the Contrast condition. This means that participants perceived a difference between the optimally informative baseline in the Contrast condition and the No-Contrast condition, which they judged to contain more information than necessary. Interestingly, no parallel effect was found for MinAAs, suggesting that participants did not perceive differences between the degree of informativity of the two conditions tested. Results also revealed that the magnitude of the effect of perceived informativity was not the same for ColAs, RelAs and MaxAAs. The results from the 2-way ANOVA interaction and the t-tests indicate that the effect was bigger for ColAs and MaxAAs than it was for RelAs, while no significant difference in perceived informativity was found between ColAs and MaxAAs. The main conclusion that can be extracted from these results is that perceived informativity is indeed modulated by adjective class. In the general discussion (§3.4), I address the relationship between perceived informativity and RECs.

3.4 General Discussion

Out of the four adjective classes tested in Experiment 1, RECs were detected for ColAs, RelAs and MaxAAs. However, MinAAs failed to display a REC, as target vs. competitor disambiguation took place in the same time window, i.e. W4, for both the Contrast and the No-Contrast condition. The

results of Experiment 2 also showed important differences between adjective classes, such that all adjective types tested gave rise to a perceived informativity penalty in the No-Contrast condition compared to the Contrast condition, with the exception of MinAAs. Taken together, Experiments 1 and 2 suggest that informativity is an important factor in RECs: adjectives that showed an overspecification penalty (ColAs, RelAs and MaxAAs) also gave rise to a REC, whereas adjectives that did not show and overspecification penalty (MinAAs) did not display a REC. However, the timing differences observed in the RECs of ColAs, RelAs and MaxAAs could not be uniquely attributed to the overspecification penalties detected by Experiment 2. As discussed above, the magnitude of the perceived informativity effect was different across these three adjective types, with RelAs showing a significantly smaller effect compared to ColAs and MaxAAs, for which the size of the effect was comparable. If perceived informativity were the only source of RECs, we would expect ColAs and MaxAAs to pattern alike with respect to the timing of their RECs and show earlier effects compared to RelAs. However, the results of Experiment 1 display a different pattern, with MaxAAs being delayed with respect to both ColAs and RelAs. Finally, the results from Experiment 2 also help understand why MinAAs did not pattern with ColAs in the eye-tracking results: the lack of an overspecification penalty suggests that MinAAs do not give rise to informativity-based reasoning, i.e. the use of an overspecified MinAA is not perceived to be redundant. Since this type of reasoning has been argued to support informativity-based RECs, the lack of a REC for MinAAs is to be expected given that the interpretation of this adjective class does not seem to be subject to an overspecification penalty.

3.4.1 Magnitude of the Perceived Informativity Effects

Even though the magnitude of the perceived informativity effects did not align with the timing of the RECs, it should be noted that the magnitude of the perceived informativity effects did correlate with the magnitude of the RECs. Results from Experiment 2 show that the effect size was the greatest for MaxAAs, whereas RelAs presented the smallest effect with ColAs showing an effect

of intermediate strength. This is the same exact pattern found in the relative strength of the three RECs found in Experiment 1. Taken as a whole, these findings suggest that overinformativity might actually be a good predictor of the strength, though not the timing, of RECs.

3.4.2 ColAs and Overinformativity

The findings pertaining to ColAs deserve more elaboration. First, given the abundance of results showing that speakers have a greater tendency to use ColAs redundantly than any other class of adjectives (see Pechmann 1989; Belke & Meyer 2002; Nadig & Sedivy 2002; Sedivy 2004, among many others), the clear penalty for overspecified uses of ColAs found in Experiment 2 is somehow surprising. If hearers are sensitive to the probabilities of use of overspecified adjectives, ColAs would be expected to give rise to the lowest overspecification penalty among all the adjectives tested in Experiment 2. It is possible that the nature of the stimuli used in our experiment had an effect on how overinformative ColAs were perceived to be. In a production experiment, Rubio-Fernández (2015) shows that the rates of overspecification of ColAs vary depending on the nature of the object. Specifically, Rubio-Fernández found lower rates of color overspecification with geometric shapes in polychrome displays compared to displays containing garments, a type of object for which color is a more central feature. Given that the production of overspecified ColAs is lower for geometric shapes, which are precisely the type of objects tested in Experiment 2, it is not unexpected that ColAs showed an overspecification penalty in Experiment 2.

Second, the results from Experiment 1 showed a clear REC for ColAs. However, in a previous study, Sedivy (2004) failed to find a REC for this same class of adjectives. The only difference between Sedivy's study and the current study is that Sedivy tested artifacts, not geometric shapes. The explanation for the incongruence between the current ColAs results and Sedivy's results might also have to do with Rubio-Fernández's finding that real world objects are more often described with redundant ColAs than geometric shapes. Furthermore, in a production study, Sedivy (2004) reports that participants produced redundant ColAs more often when describing real world artifacts

than other classes of adjectives such as material adjectives (e.g. *wooden*) or RelAs. Taken together, these findings suggest that ColAs give rise to different overspecification rates depending on the type of object being described by the adjective. The clear prediction that arises is that there should exist an inverse relation between the overspecification rate found in production data (which in its turn is contingent on the type of objects under consideration) and perceived informativity measures such as the one used in Experiment 2. If this prediction is on the right track—and provided that informativity-based RECs are only triggered by adjective/noun combinations that give rise to low overspecification rates in production—, the apparent discrepancy between Sedivy's results and the results from Experiment 1 would be explained away.

3.5 Conclusion

This chapter reports results from two experiments investigating the patterns of context-sensitivity of GAs (i.e., RelAs, MaxAAs and MinAAs). Following Sedivy et al.'s (1999) experimental design, Experiment 1 consisted of a Visual World eye-tracking study that had the goal of assessing whether information from the visual context was integrated comparably during the processing of RelAs, AAs and ColAs. Results show that ColAs display a different sort of context-sensitivity compared to both RelAs and AAs. In particular, I have argued that the REC observed for ColAs was driven by pragmatic reasoning about the Gricean maxims of quantity and manner, whereas other sources of context-sensitivity played a role in deriving the RECs associated with RelAs and MaxAAs. The current results constitute the first explicit attempt at characterizing the processing signature of RECs that are shaped by Gricean pragmatic reasoning vs. RECs that are further shaped by other types of interactions between meaning and context. More specifically, I have proposed that the processing signature distinguishing these two types of RECs could lie in the role of the competitor object, i.e., whether the REC is driven by an increase of fixations to the target object, or a decrease in fixations to the competitor object, and whether the competitor object is considered a privileged referent candidate along with the target before disambiguation between these two objects takes

place.

The second goal of Experiment 1 was to provide new empirical evidence to tease apart theories that assume that the lexical thresholds of AAs are context-insensitive, from theories that model AA threshold as free variables that can take any value in the adjectival scale as a function of the context of utterance. To approach this question, I hypothesized that theories assuming shiftable thresholds for all GAs would predict that RelAs and AAs should behave similarly with respect to the way in which contextual information is exploited during online processing, since all gradable thresholds would be fixed by resorting to contextual information. On the other hand, theories that assume variable thresholds for RelAs and fixed thresholds for AAs predicted a more complicated pattern of results. Specifically, it was hypothesized that Experiment 1 should show a three-way distinction among GAs because RelAs are assumed to be lexically context-sensitive, whereas MaxAAs are only pragmatically context-sensitive (i.e. they are subject to imprecision calculation). Finally, MinAAs were argued to not be subject to either of these two types of interactions between meaning and context. The results of Experiment 1 patterned in line with theories that pose fixed gradable thresholds. Both RelAs and MaxAAs displayed sensitivity to context in the form of RECs, but the timing of the effects was very different. For RelAs, the REC took place in the adjective window, whereas for MaxAAs the effect took place at a point where participants were processing the head noun and possibly integrating the information of the full modified NP. I have argued that the fact that MaxAAs only showed contextual effects in the noun window, not in the adjective window, is consistent with the view that MaxAAs only interact with context to set the value of a pragmatic precision threshold. Finally, Experiment 1 also revealed that the processing of MinAAs was not sensitive to the properties of the visual context, as no REC was detected for this adjective class. MinAAs constitute perhaps the best testing area to address the status of absolute thresholds, as the two types of theories considered in this chapter make clearly different predictions with respect to Experiment 1. The fact that information about context seemed to not play a role in the processing of MinAAS constitutes an important piece of evidence in favor of the view that absolute thresholds

are not set via context.

Finally, Experiment 2 was conducted in order to control for the possibility that the processing differences attested among GAs in Experiment 1 resulted from differences in the strength of the expectations about contrastive use associated with each of the four adjective types tested in the experiment. With this goal in mind, Experiment 2 was designed to quantify the magnitude of such expectations for both ColAs and GAs. The results show that while Gricean reasoning about quantity and manner is definitely one of the drivers of RECs, it cannot, by itself, account for the timing differences between RelAs and MaxAAs observed in Experiment 1, thus reinforcing the conclusion that other grammatical and pragmatic factors played a role in deriving the RECs uncovered by the experiment. Furthermore, Experiment 2 also explains why MinAAs did not pattern with ColAs in giving rise to a Gricean REC, since MinAAs were the only adjective class that was not associated with a default contrastive interpretation. To conclude, the results of Experiments 1 and 2 argue against a homogeneous theory of GAs when it comes to the role that context plays in their interpretation, and argue for a more nuanced view that leaves room for variability both in the semantic and pragmatic representation of GAs.

CHAPTER 4

CONTEXT-SENSITIVE INTERPRETATIONS OF NUMERALS

4.1 Introduction

Numerals give rise to at least three types of readings exemplified in (45). The lower-bound or *at least* interpretation of numerals (45a) has been object of much investigation. Its status, i.e. whether it is part of the core meaning of numerals, o whether it is derived by some semantic or pragmatic mechanism, remains a question of much debate and will not be addressed in detail in the current chapter.[1] Rather, this section focuses instead on the latter two readings of number words (45b-c), usually referred to as the *precise* and *imprecise* interpretations of the numeral. In (45b), the numeral receives a doubly-bound interpretation and is therefore construed as a precise measure (45b), whereas in (45c) the numeral is unbound, and construed as a fuzzy range (45c).

(45) Mar has *60* publications.

a. Lower-bound interpretation: Mar has *at least* 60 publications.

b. Doubly-bound interpretation: Mar has *no more and no less than* 60 publications.

c. Unbound interpretation: Mar has *approximately* 60 publications.

In Krifka's granularity theory, introduced in Chapter 2, readings (45b) and (45c) can be modeled using different granularity scales. Coarser-grained scales result in more approximate interpretations because less granules of information are represented in such scales compared to finer-grained ones, see Figure 4.1. For instance, if scale d in Figure 4.1 was adopted for the interpretation of sentence (45), the numeral *60* could receive any value in the range spanning from 56 to 64. If, on the other hand, the numeral was interpreted with respect to the coarser-grained scale c, it would receive a more approximate interpretation, as the numeral could be assigned any value in the range spanning from 46 to 74.

1. Although see §4.3 for some cursory discussion and representative references of the different proposals.

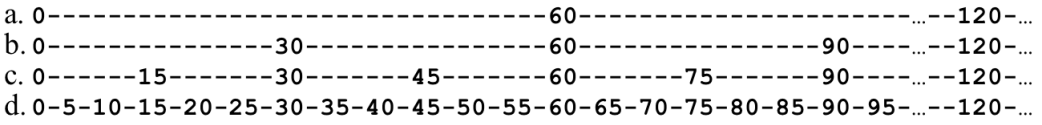


Figure 4.1: Four possible granularity scales (a-d) of the number line starting at 0 and followed by the positive integers. Image reproduced from Krifka (2007).

In the remainder of this chapter, I discuss context-sensitive interpretations of numerals such as (45c). I start by describing the properties of unbound interpretations of numerals in §4.2, where I review the parallels between the context-sensitivity displayed by numerals and MaxAAs. In §4.3, I discuss the phenomenon of *slack regulation*, i.e. the use of linguistic operators such as *exactly* or *approximately* to overtly signal the intended precision threshold to the interpreter. Section 4.4 discusses an observation originally made by Krifka (2007), namely that interpreters default to the more imprecise interpretation tolerated by the context. I introduce some functional explanations of this heuristic articulated around the assumptions that: a) linguistic communication is rational; and b) favoring an imprecise interpretation over a precise one constitutes the rational choice under uncertainty. Section 4.5 introduces the notion of *cost*, a concept that underlies all accounts of the preference for imprecision. Among the different types of cost discussed in this section is the idea of processing cost for the comprehender. This type of cost will be the center of the experiments presented in Chapter 5. Finally, in section 4.5.1 I discuss the few precedents that have investigated imprecision from a processing perspective. In §4.6, I conclude.

4.2 Context-sensitivity of numerals is *imprecision*

In this section, I describe the properties of context-sensitive interpretations of numerals, and point out that the patterns of context-sensitivity observed for number words align with those of AAs, and diverge from those shown by RelAs. Based on this, it is argued that variable interpretations of numerals should be treated as an instance of imprecision.

The most salient restriction governing context sensitive uses of number words is that some numerals lend themselves to variable interpretations more easily than others. In particular, round numbers (for our purposes integers ending in zero) can easily be construed as a fuzzy unbound range, while non-round numerals resist such an interpretation. This is illustrated by the contrast in (46): round numbers, which tolerate approximate interpretations, can later be precisified, as shown by the felicitous continuation in (46a), while non-round numbers, which tend to receive a default precise interpretation, cannot. The continuation in (46b) cannot be understood as a precisification, only as a contradiction. Provided that the first numeral already received a precise interpretation, the continuation is simply incompatible with the speaker's previous statement. On the contrary, by uttering (46a) the speaker is not being contradictory. Her continuation is understood as a refinement of her previous approximate statement.

(46) a. 5,000 people live in Alaró. Actually, to be precise, 5,004 people live in Alaró.

b. 5,003 people live in Alaró. #Actually, to be precise, 5,004 people live in Alaró.

Example (46a) shows that an unbound interpretation of a number word can always be precisified. However, the opposite is much harder to accomplish: precise numerals do not easily undergo *im*precisification, as shown by the asymmetry in the acceptability of the discourses in (47b) and (48b).

(47) a. Biel: 5,000 people live in Alaró.

b. Fer: Actually, exactly 5,004 people live in Alaró.

c. Biel: Well, yeah, ok.

(48) a. Biel: 5,004 people live in Alaró.

b. Fer: #Actually, approximately 5,000 people live in Alaró.

c. Biel: #Well, yeah, ok.

As seen in (48b), the unbound interpretation of a round number cannot be used to challenge the precisely interpreted numeral in (48a). Once a more informative numeral has been provided (i.e.

the precisely interpreted number), it is not possible to switch back to to a less informative precision threshold, i.e. one that would license the use of imprecise numerals. This property is often referred to as *unidirectionality* (Lewis 1979; Klecha 2014, 2017), and basically captures the fact that while raising the standard of precision is a discursively legal move, lowering the standard of precision is a lot harder to accomplish. To put it differently, cases of metalinguistic disagreement (Barker 2002; Klecha 2017) such as (47)-(48), where the value of some linguistic variable—in this case the precision threshold—is under negotiatioation can only be resolved in one direction: speakers can agree to raise the standard of precision effortlessly by means of conversational moves such us the ones exemplified in (47). However, the opposite, i.e. lowering the standard, while not impossible requires a much more explicit kind of metalinguistic discussion (see Klecha 2017 for specific examples).

As was discussed in Chapter 2, AAs behave like numerals in that they can also be precisified. RelAs, on the other hand, cannot undergo precisification. Therefore, with respect to precisification and *unidirectionality*, numbers pattern like AAs (51)-(52) and unlike RelAs (49)-(50).

(49) a. Biel: That table is big.

b. Fer: No, it's not big!

c. Biel: Yeah, ok I guess it is not big.

(50) a. Biel: That table is not big.

b. Fer: No, it is big!

c. Biel: Yeah, ok I guess it is big.

In (49) and (50), Fer and Biel can converge on a new threshold value for the adjective *big*, regardless of whether this new value ends up being higher (49) or lower (50) than what Biel had initially in mind. However, much like what we saw with numerals, AAs display more stubborn precision thresholds, because they resist being lowered, as shown by the contrast in the acceptabilty

of Fer's remarks in examples (51) and (52).[2]

(51) Biel and Fer work for a dentist and are discussing whether they can schedule one more emergency appointment for the day after. For that day, the schedule is already filled with back to back appointments with the exception of one free time slot from 8 to 8:30 AM, which usually never gets filled.

a. Biel: The agenda is full, we cannot schedule any more appointments.

b. Fer: No, it's not full!

c. Biel: Well, yeah, ok, I guess we can squeeze in one more person at 8 AM.

(52) Same context as (51).

a. Biel: The agenda is not full, we can schedule one more appointment.

b. Fer: #No, it is full!

c. Biel: #Well, yeah, ok, I guess we cannot squeeze in one more person at 8 AM.

From the discussion above, it seems uncontroversial that numerals display a type of context-sensitivity that parallels that of AAs, not that of RelAs. Based on this, in the remainder of this dissertation, I will refer to context-sensitive uses of numerals as instances of imprecision.

4.3 Slack Regulation

Speakers can choose to linguistically encode the precision threshold with which they intend a numeral to be interpreted:

(53) a. {*Exactly*, *not more and not less than*} 100 investors acquired stock.

b. {*Approximately*, *about*} 100 investors acquired stock.

The modifiers in (53a) have the effect of rising the precision threshold, whereas the modifiers in (53b) indicate that the numeral following them should be interpreted imprecisely (Lakoff 1973;

2. See Klecha (2017) for further discussion.

Sadock 1977; Morzycki 2012; Anderson 2014; Ferson *et al.* 2015; Anderson 2015, among others). Lasersohn (1999) treats modifiers that have the effect of increasing the precision level as identity functions that, despite not having any truth-conditional effects, make a pragmatic constribution, namely shrinking the size of the halo of the expression they modify.

Lasersohn takes the status of the modifiers in (53b) to be fundamentally different. Based on examples like (54), it is argued that these expressions do not directly regulate the pragmatic halo, but rather have the function of equating the semantic truth-conditions of the expression to the pragmatic halo denotation. Lasersohn also proposes that the original semantic content of the expression is subtracted from the final denotation, following the intuition that sentence (54) would be false in case John were indeed the king. Therefore, in this view, slack tighteners like *exactly* are pragmatic operators, while hedges like *loosely speaking* are semantic operators.

(54) *Loosely speaking* John is king.

Lasersohn's views are not however without counterexamples. Lasersohn himself notices that his treatment of slack tighteners makes the wrong predictions for numerals, since in cases like (55) the Slack Regulator *exactly* seems to interact with the semantic denotation of the numeral.

(55) a. John has 3 homes, in fact he owns 5.

b. #John has exactly 3 homes, in fact he owns 5.

Second, it is not clear that all hedges have the effect of excluding the semantic—for our purposes *precise*—denotation of the expression susceptible to be interpreted imprecisely. If this was the case, it is predicted that an imprecise interpretation of a number word induced by a hedge should not tolerate precisification. While this might be the case for propositional hedges like *loosely speaking* (56a), scalar hedges like *about* do not behave the same way, as seen in (56b).

(56) a. *Loosely speaking*100 investors acquired stock. #In fact, exactly 100 investors acquired stock.

b. *About*100 investors acquired stock. In fact, exactly 100 investors acquired stock.

Given that the distinctions between hedges and slack regulators proposed by Lasersohn do not seem to be robust, in this dissertation I will assume the null hypothesis, namely that slack regulators form a natural class and should receive a unified analysis. This analytical option has actually been developed by Sauerland & Stateva (2007), who treat approximators such as *exactly* or *approximately* as granularity functions that either select the finest or the coarsest granularity level available in the context for the evaluation of their complement.

4.4 A bias towards imprecision?

It has been argued that comprehenders are biased to interpret round numbers approximately (Krifka 2002, 2007; Bastiaanse 2011; Jäger 2012; Kao *et al.* 2014; Solt *et al.* 2017; Klecha 2017). Krifka (2002) observes that short expressions tend to correlate with approximate interpretations, whereas longer expressions tend to correlate with more precise interpretations. For number words, this difference roughly aligns with the round vs. non-round distinction, as the former tend to be shorter than the latter. For instance, the non-round number *fourty-nine* is longer in number of syllables than the round number *fourty*. Krifka captures this empirical generalization with the following principle:

(57) Round Number/Round Interpretation (RN/RI) principle:

(57) Short, simple numbers suggest low precision levels.

(58) Long, complex numbers suggest high precision levels.

In later work, Krifka (2009) qualifies this descriptive generalization to accommodate apparent counterexamples such as (59-61):

(59) a. Mary waited for forty-five minutes. (approximate)

b. Mary waited for forty minutes. (precise)

(60) a. The wheel turned one hundred and eighty degrees. (approximate)

b. The wheel turned two hundred degrees. (precise)

(61) a. Her child is eighteen months. (approximate)

b. Her child is twenty months. (precise)

All the numerals involved in the a-examples above tolerate imprecise interpretations, despite being longer in number of syllables than the numerals used in the b-examples, which are claimed to only receive a precise interpretation. Based on this, Krifka argues that the bias towards imprecise interpretations cannot be uniquely driven by the simplicity of the linguistic expressions themselves. Krifka proposes that simplicity of *representations* is also a strong predictor of when a hearer should opt for an imprecise interpretation. For instance, in example (59) 'fourty-five minutes' corresponds to three-quarters of an hour, in (60) 'one hundred and eighty degrees' is half of a circumference and 'one hundred and eighty degrees' in (61) corresponds to a year and a half. Thus, what (59-61) show is that the scale in which the numerical expression appears is an important factor: more conceptually salient units in the scale will be more prone to be interpreted imprecisely. Krifka goes on to point out that the bias towards imprecise interpretations can be better understood as a bias for coarse-grained representations, as stated in (62).

(62) *Coarsest Scale Principle* (CSP):

If a measure expression α occurs on scales that differ in granularity, then uttering α implicates that the most coarse-grained scale on which α occurs is used.

The CSP therefore constitutes a heuristic for hearers to determine when they should favor an imprecise interpretation over a precise one. An independent matter is how to represent the actual granules of the coarsest-grained scale in a given context. This is a question that probably cannot be answered by uniquely looking at the semantics of number words, or the words they modify. It is very likely that questions related to number cognition (e.g. what the most accurate representation of the mental number line is—i.e. whether it is linear or logarithmic—, or how the approximate

and exact number systems relate to each other in healthy adults[3] constrain the range of logical possibilities for plausible granularity representations.[4]

If comprehenders follow interpretational biases such as the CSP, it is then unclear why speakers would choose to explicitly signal a low precision standard through an approximator.[5] I see at least two possible answers to this question. The first one is that hedges such as *approximately* might mostly be used in discursive contexts that already presuppose a high precision standard. As discussed in §4, lowering the precision threshold is a much cumbersome discursive move than raising it. By using a hedge, a speaker can smoothly and uncontroversially lower the threshold of precision, without the need of engaging in a metalinguistic negotiation with the addressee in order to converge on a precision threshold value. If on the other hand hedges can be felicitously used in any precision context, which I suspect is the case, the use of a hedge can still be informative, even if interpreters follow the CSP heuristic. As stated, the CSP does not operate on the value of the precision threshold, at least not in the sense that slack regulators do. The CSP simply guides the interpreters' choice of granularity scale *given* the scales made available by the contextual precision threshold. Given this distinction, the use of a hedge could be just seen as a way to ensure that an imprecise interpretation obtains in contexts where the speaker has reasons to suspect that the interpreter might otherwise adopt a too precise standard, and therefore a too precise interpretation, even after having followed the CSP.

The CSP, or more generally the preference for imprecise interpretations over more precise ones, has received functional accounts showing that from a probabilistic perspective the choice of imprecision over precision is the rational choice for the comprehender, at least when there is complete uncertainty about the imprecision threshold, i.e. when the comprehender takes the chances of the imprecise and the precise interpretation to be equiprobable. To illustrate this point, I reproduce

3. See Dehaene (2003); Izard & Dehaene (2008); Izard *et al.* (2008), among many others.

4. This is at the moment a very open question that is beyond the scope of this dissertation, but see Hobbs & Kreinovich (2006); Cummins *et al.* (2012); Lefort *et al.* (2017) or Cummins (2018) for some preliminary results.

5. I thank Sally McConnell-Ginet for bringing this question to my attention.

Krifka's argumentation following Klecha (2017). Suppose John asks Katie how many people attended her keynote lecture and Katie answers that *100* people went to the lecture. Assume that John knew that the maximum capacity of the lecture hall was 200 people, and that he took any number between 1 and 200 to be equally as likely. If John assumes a precise interpretation of Katie's utterance, there is a 1 in 200 chance of *no more and no less than 100* being the true answer to his question answer. Given that John is completely uncertain about the precision level (which here I assume, for simplicity, is discrete, i.e. either precise or imprecise), there is a 0.25% probability that John assumes a high standard and the value of Katie's utterance is *no more and no less than 100*. If on the other hand, John interprets Katie's answer as a range, meaning for example between 90 and 110, and assumes a low standard, the chances that Katie's answer was in that range is of 10.5%. Therefore, the probability of a low standard given Katie's answer is more than 97.5%. To put it differently, John has better chances of correctly inferring Kate's intended message if he assumes a weaker message that is compatible with multiple values of the numeral, even if it is at the cost of assuming a less informative interpretation.[6]

4.5 Cost

Effort minimization (and therefore cost minimization) has been claimed to be an important principle of human behavior Zipf (1949). In the Gricean pragmatics tradition, cost minimization is the logic driving maxims such as Quantity and Manner, both of which have the goal of optimizing the amount of information, usually correlated with the number of words, transmitted by the speaker.[7] When it

6. See also Bastiaanse (2011) for a similar argument cast in game theoretic terms, and Kao *et al.* (2014) for an account of imprecision within the Rational Speech Act framework. For similar arguments applied to the domain of vagueness se Frazee & Beaver (2010) and De Jaegher & van Rooij (2009).

7. Neo-Gricean theories also capture the speaker's tendency to be economical with respect to her utterances with the R-Principle or the I-Principle proposed by Horn (1984) and Levinson (2000) respectively, which state that speakers should say as little as possible as long as they can be understood (i.e. as long as the hearer has enough grounds to infer the intended communicated meaning).

comes to expressing numerals, such pressures might lead speakers to use rounder numbers, which as discussed above are often shorter than their neighboring non-round numbers, e.g. choosing to say *fifty* instead of a more precise but longer representation of the intended meaning, e.g. *fifty-one*. This sort of production pressures might be one of the underlying causes that have led round numbers to be associated with approximate interpretations and not the other way around.

The view that imprecision is at least partially driven by economy of production considerations is not necessarily extensible to the comprehender. When confronted with a bare round number, the comprehender must perform a complex inferential task: first, she must determine the threshold of precision. Second, in case of assuming a low threshold, she must decide a plausible range of interpretation for the numeral. Therefore, while imprecision might be a good way of optimizing cost from the speaker's perspective, it is conceivable that imprecision is costlier from for the comprehender. In this respect, heuristics such as the CSP might be seen as a strategy developed by comprehenders to cope with the uncertainty involving the interpretation of bare numerals.

4.5.1 Processing imprecision

In this section, I consider in more depth the hypothesis that imprecise meanings are easier to process for the hearer. For instance, Krifka (2007; 2009) suggests that 'more coarse-grained representations of information might be cognitively less costly than a more fine fine-grained one' (p. 2). However intuitive this claim is, there barely exists any research bearing on this question. One exception is work done within Relevance Theory (Sperber & Wilson 1995; Wilson & Sperber 2002; Sperber & Wilson 1985). In this framework, relevance is taken to be a function of cognitive effect and processing cost: relevance is maximized when the speaker chooses an utterance that minimizes the processing cost for the speaker while still being informative. In some cases, speakers aim at maximizing relevance, even at the cost of truthfulness; a relevant utterance is not necessarily one whose propositional content is true, but rather one that entitles hearers to derive true conclusions from its interpretation. Applied to the case of imprecision, this means that speakers will choose to

be imprecise when the cognitive effects of the imprecise utterance (defined as the true consequences the hearer is entitled to derive from the utterance) are indistinguishable from those of its precise counterpart. van der Henst *et al.* (2002) provide evidence from a field experiment that is compatible with this view. van der Hest and colleagues conducted a study where participants were requested to provide an answer to the question '*what time is it?*'. The authors found that there was a significant tendency for participants to provide answers that contained round times (for the purpose of this experiment, only times ending in a multiple of five minutes were considered *round*). Importantly, this was also the case for the group of participants that was carrying a digital watch, suggesting that the choice of a round time could not solely be attributed to some degree of epistemic uncertainty about what the exact time was. While it is possible that participants carrying an analogue watch were to some extent uncertain about the exact time on their watches, this could have not been the case for those participants who carried a digital watch.

Interestingly, van der Hest and colleagues found a different pattern of results when the confederate researcher changed the request to '*Hello! My watch is going wrong. Do you have the time please?*', a request that suggested that the researcher needed an exact time in order to set his own watch. The results showed that the proportion of round times was significantly lower in the experimental group, compared to a control group who provided the time as an answer to the question '*Do you have the time please?*' (see also Gibbs & Bryant (2008), who found comparable results using a very similar paradigm). The authors interpreted these results as evidence that, in planning their utterances, speakers seek to maximize relevance to their audience.

Even though van der Hest et al. (2002) do not mention this point explicitly, the logic of their argument is built upon the assumption that imprecise utterances are easier to process compared to utterances that are only compatible with a precise interpretation. Unfortunately, no evidence is provided to support such an assumption (see also Solt *et al.* (2017) for a similar point).

In recent work, Solt *et al.* (2017) address a somewhat related question. The authors investigate whether there exists a processing advantage for rounder numerical expressions and whether this

advantage is driven by the degree of roundness, or by the granularity of the scale in which the number occurs. Solt *et al.* (2017) tested these hypothesis experimentally using a short-term memory task in which participants saw sequences of 3, 4 or 5 clock times. Participants were told to interpret these sequences of times as departure times for trains. Each of these sequences was followed by a probe time. The task consisted of deciding whether the target time occurred in the sequence created by the previous clock times or not. Three granularity levels (coarse, medium and fine) were tested (63):

(63) Granularity Scales Tested by Solt *et al.* (2017).

a. Coarse (15-minute granularity): –:15, :30, :45

b. Medium (5-minute granularity): –:10, :25, :40

c. Fine (1-minute granularity): –:21, :36, :51

While both the coarse and the medium granularity levels contained round numbers, the fine granularity level was formed by clock times that could not be considered round. Example (64) contains one of the experimental items tested, where all the target times were compatible with the granularity scale of the clock times in the sequence:

(64) Item example (3 clock-time sequence) used by Solt *et al.* (2017).

a. Coarse: 2:45, 6:30, 8:45 Probe: 2:30

b. Medium: 2:10, 6:25, 8:40 Probe: 2:25

c. Fine: 2:36, 6:51, 8:21 Probe: 2:51

The results showed that participants' accuracy significantly declined as the length of the clock-times sequence increased. Also, participants were significantly less accurate for fine vs. coarse granularity scales, although no significant difference was found between coarse and medium granularity. Finally, a significant interaction between sequence length and fine granularity was also found, such that participants performed similarly in the fine and coarse granularity levels at sequence length 5. With respect to Reaction Times, participants took significantly longer to respond

for fine vs. coarse-grained granularity. However, similar to what was observed for the accuracy results, no significant difference was attested for medium and coarse granularity. The fact that differences between granularity levels only emerged for fine vs. medium/coarse granularity levels suggests that roundness is a better predictor of the results than scale of granularity.

These results should be interpreted with caution, as it is not necessarily the case that participants needed to infer a scale of granularity in order to successfully perform the experimental task. Note that in the item example in (64), the digits corresponding to the minutes in the target time are identical to the minutes in one of the clock times that formed the sequence. Therefore, it is possible that participants simply developed the strategy of memorizing the two digits corresponding to the time minutes in the sequence and uniquely relied on that information in order to decide whether the probe time belonged to the sequence or not. In other words, it is possible that participants never used the clock times in the sequence to infer the granularity scale. This might be the reason why no clear effects of granularity were found in the first place. Given that previous research has found that rounder numbers are easier to recall than non-round numbers (Mason *et al.* 1996), the patterns of results reported by Solt et al. could be alternatively explained as a result of different limitations in short-term memory for round vs. non-round numbers rather than as the result of granularity calculation. Therefore, Solt et al.'s results do not rule out the hypothesis that scale of granularity plays a role in the processing of numerals. More importantly, their experimental manipulations did not control for whether participants interpreted the clock times imprecisely in the coarser-grained granularity conditions. It is possible that a different task that 1) truly forced participants to compute the scale of granularity; and 2) ensured that participants adopt an imprecise interpretation of the numeral in the relevant conditions would be able to detect a processing advantage for numbers that occur in coarser-grained scales, compared to numbers that only occur in finer grained ones.

An open question remains of why there exists a processing asymmetry between round and non-round numbers, and whether these differences in processing are linked to imprecision in the first place. After all, the only numbers that tolerate imprecise interpretations are round numbers.

Solt et al.'s study cannot speak to this latter question because all the numbers used in their study were most likely interpreted precisely (recall that participants were instructed to interpret the clock times as train departure times). Regarding the first question, Solt et al. suggest that the processing advantage for round numbers might result from the fact that crosslinguistically round numbers are more frequently used than non-round numbers (Sigurd 1988; Dehaene & Mehler 1992; Jansen & Pollmann 2001). It is also a well-known fact that frequency plays a role in processing. For instance, there exists as a big body of research showing that word recognition takes place faster for high-frequency words, compared to low-frequency words (Marslen-Wilson & Welsh 1978; Luce & Pisoni 1998; Dahan *et al.* 2001, among many others). Whether these facts about frequency are the cause of the correlation between round numbers and imprecise interpretations, or whether higher frequency is a consequence of this correlation is hard to tell. On the one hand, it is possible that imprecise interpretations are more readily available for high-frequency words. However, a second possibility is that round numbers are more frequent precisely because they support approximate interpretations, which would license their use in more contexts thus increasing their frequency.

Be it as it may, the hypothesis about a potential link between imprecision and ease of processing makes the clear prediction that we should be able to observe differences in processing of the *same* round number when interpreted precisely vs. imprecisely, such that imprecise interpretations should show less cost. In fact, it is also predicted that a similar processing advantage should be expected to arise for any lexical expression that could give rise to (im)precise interpretations. To date we do not have any behavioral evidence to confirm or falsify this hypothesis. The sequence of experiments presented Chapter 5 is aimed at addressing this precise question.

4.6 Conclusion

In one way or another, most work on imprecision assumes an asymmetry between precision and imprecision such that the former is costlier than the latter. However, the nature of this cost is usually not fleshed out. In the following chapter, I explore one possible option, namely that the

asymmetry between precision and imprecision arises from a difference in processing cost for the comprehender, such that imprecise meanings are easier to process for the hearer than their precise counterparts.

CHAPTER 5

PROCESSING (IM)PRECISE INTERPRETATIONS OF NUMERALS

5.1 Introduction

The current chapter presents a series of experiments whose goal is to address the second research question introduced in Chapter 1:

(65) Are imprecise representations of number words less costly to process for the comprehender?

To tackle this question, I conducted a series of self-paced reading studies, where I examine the processing of round numerals (for the purpose of this dissertation, I take round numerals to be numerals ending in 0) when interpreted precisely vs. imprecisely. In this respect, the experiments presented below differ from previous studies in that I compare precise vs. imprecise representations of the *same* numeral, as opposed to comparing round vs. non-round numbers. This approach has the advantage of focusing on a true minimal pair, i.e. two representations that differ uniquely with respect to the precision level with which they are interpreted, allowing to control for potential confounds introduced by the round vs. non-round distinction.

The studies investigate the processing of (im)precise interpretations that come about via two different mechanisms: 1) through the inclusion of a slack regulator that explicitly signals the precision level intended by the speaker; 2) through contextual cues about the conversational goals that the speaker had in mind when uttering the numeral. The objective of these manipulations is to determine whether different means of signaling (im)precision result in comparable processing patterns. In conditions involving a slack regulator, the precision threshold is made explicit. Therefore, from the comprehender's perspective, resolving the precision threshold—which ultimately determines what interpretation of the numeral should be favored by the comprehender—involves less uncertainty in the presence of a slack regulator. On the other hand, when no slack regulator

is present, the comprehender must uniquely rely on indirect contextual clues to *infer* the speaker's intentions and determine whether to adopt a precise or an imprecise interpretation of the numeral.

For each experiment, high vs. low precision interpretations of round numerals were compared. The results converge in showing that when comprehenders are confident about the precision threshold intended by the speaker, the processing of imprecise representations of number words systematically incurs a higher processing cost. Taken together, these results show that the claim that imprecision is less costly for the comprehender than precision is not empirically supported, at least when reading times are used as the dependent measure. I discuss these results in further detail in §5.8.

The structure of the chapter is as follows. In §5.1.1, I outline the general predictions made by the *Coarsest Scale Principle* with respect to the online processing of round numerals. Experiments 3a (§5.2), 3b (§5.3) and 4 (§5.4) explore the processing differences between precise and imprecise interpretations of number words when the precision level is set by a slack regulator. Experiments 3a and 3b test the slack regulators *approximately* vs. *exactly*, whereas Experiment 4 tests the slack regulators *about* vs. *exactly*. Experiment 5 (§5.5), which consists of an offline judgement task, closes the first half of this sequence of experiments by further examining the preferential interpretations associated with the control condition used in studies 3 and 4. Finally, Experiments 6a (§5.6) and 6b (§5.7) address the processing of (im)precise interpretations of number words in those cases where the comprehender resolves the precision level by relying uniquely on pragmatic cues related to conversational goals.

5.1.1 General Predictions

If heuristics such as the *Coarsest Scale Principle* guide comprehenders' interpretive choices of round numbers, such that there exists a bias towards adopting the most imprecise interpretation tolerated in the context, Experiments 3-6 should uncover a processing facilitation for round numbers in contexts that allow for imprecision compared to contexts that do not. If, on the contrary, there

exists a preference for precision, we should observe a processing facilitation for precise interpretations of number words over imprecise ones.

5.2 Experiment 3a

Experiment 3a was conducted to determine if the processing signature of round numbers—measured in Reading Times (RTs)—differs when the numeral is interpreted precisely vs. imprecisely. In particular, Experiment 3a investigates the effect of (im)precision on the processing of round numbers when the precision threshold is linguistically signaled by a slack regulator. The experiment uses the slack regulator *approximately* to induce an imprecise interpretation of the numeral, and the slack regulator *exactly* to bias participants towards the precise interpretation of the numeral.

5.2.1 Methods

Participants

Participants were 45 native speakers of American English between the ages of 18 and 35 (25 females, mean 21.7, range 18-35). Participants were either payed $10 for their participation, or were compensated with research credits. All subjects were naïve to the goal of the study.

Apparatus

The experiment was carried in a Dell Precision T5600 computer in a lab setting. The experiment was coded using the online platform Ibex Farm.

Procedure

Participants sat in front of a computer screen, where they received instructions from the experimenter before starting the study. Immediately after receiving the instructions, participants initiated

the practice session where they performed three practice trials that helped them familiarize with the experimental task and ensured that they had understood the instructions. At that point, participants proceeded to the experiment. For each trial, participants first saw a context paragraph that they were instructed to read carefully. After the context paragraph, the next screen presented the target sentence in a self-paced, word-by-word fashion. The target sentence screen initially showed a series of underscores as a sentence outline. Each of the underlines was replaced by a word one at a time, and reverted back to underlines, as participants tapped the space bar to move from one word to the next in the sentence. Sentences that were too long to be presented on one line were broken into two lines. After the target sentence, participants where asked a comprehension question about the immediate preceding trial. The questions always targeted information that had been provided in the contexts and in no cases involved judgements about the numeral (see the item example in (66) for a comprehension question example). Three possible answers to the question were provided. The order of the answers was randomized from trial to trial. Participants selected the answer that they believed to be correct by clicking on it with the mouse. The reading times of the experimental and filler target sentences were recorded. Response times and accuracy of the answers provided by the participants were also registered. The experiment lasted approximately 30 minutes.

Design and Materials

The study employed a fully crossed 3x3 design where the factors ADVERB TYPE and NUMBER SIZE were manipulated. The factor ADVERB TYPE consisted of three levels (*Approximately* vs. *Exactly* vs. *None*) labeled after the adverb used to induce the intended precision level. The condition with no adverb was conceived as a control representing the default interpretation of the numeral given the previous linguistic context. The second factor, NUMBER SIZE, manipulated the size of the round number included in the target sentence and consisted of the three levels *Small* vs. *Medium* vs. *Big*. Within each item, the three number sizes were constructed by adding a 0 to the right of the numeral, with the smallest number tested among the Small Numbers being 10 and the biggest number tested

among the Big Numbers being 90,000. The 9 conditions resulting from this full factorial design are exemplified in (66).

(66) a. **Context:** When the doctor arrived that morning she asked the nurse {**approximately/ exactly/** ∅} how many patients were waitlisted for the new clinical study.

b. **Target Sentence:** The nurse replied that {10/ 100/ 1,000} patients were currently waitlisted.

c. **Comprehension Question:** When did the conversation between the nurse and the doctor take place?

d. **Possible Answers:** 1) Morning; 2) Night; 3) Afternoon.

All experimental items consisted of a linguistic context (66a) followed by the target sentence (66b) containing the numeral. In order to control for sentence final wrap up effects, the noun region, which is one of the regions of interest along with the numeral region, never coincided with the end of the sentence and was always followed by at least three more words. A total of 27 items were constructed. Furthermore, 33 filler sentences were included. Fillers were of 3 types. Eleven of the fillers contained the adverbs used in the experimental items (i.e. *exactly* and *approximately*) but did not contain a number in the target sentence. The remaining fillers (a total of 22) contained adverbs ending with the suffix *-ly*, such as *definitely* or *completely*, in the context. Eleven of these fillers contained a number in the target sentence, whereas the remaining 11 did not. The 33 experimental items were divided into 9 lists following a Latin Square design. All participants saw the same set of 33 fillers described above.

5.2.2 Predictions

If imprecise interpretations of round numbers involve less processing cost than their precise counterparts, a processing facilitation of the round number and/or subsequent regions is expected for the

approximately-condition compared to the *exactly*-condition. If bare round numbers tend to be interpreted imprecisely, as stated by the *Coarsest Scale Principle* the control condition, which in this study represents the default interpretation of the bare numeral, should patterns more similarly to the *approximately*-condition than to the *exactly*-condition. Furthermore, if number size modulates the availability of an imprecise interpretation, such that bigger numbers are more likely to be interpreted imprecisely due to the fact that they are represented in more granularity scales than smaller numbers (see Figure 4.1), an interaction between approximation and number size is predicted, such that the imprecision advantage should be larger for bigger numbers.

On the other hand, if precise interpretations are favored over imprecise ones, the latter should give rise to higher reading times in the number and/or postcritical regions. In the context of the current experiment, this translates into higher reading times in the *approximately*-condition compared to the *exactly* and the no-adverb condition.

5.2.3 Results

The results reported in this section and in the remainder of this chapter were obtained from linear mixed effects regression models (Baayen *et al.* 2008) using the lmer() function from the R lmerTest package. All models were fitted with the maximal random effects structure (random by-participant intercepts and random by-participant slopes for both main effects and the interaction terms) in order to determine the maximal random effects structure supported by the data (Barr *et al.* 2013). Unless otherwise specified, all categorical predictors were treatment coded using the *No-adverb* level of the predictor ADVERB TYPE and the level *Small Numbers* from the predictor NUMBER SIZE as baselines for comparison. All main effects and interactions reported below were derived using the likelihood ratio test obtained from nested model comparison (i.e. by comparing a model containing the predictor of interest with a minimally different model that did not contain that predictor). When interaction terms in the regression models were significant, the simple slopes associated with each level of the relevant predictor were assessed by fitting the regression model separately for each

level.

Accuracy Ratings

I first analyze the accuracy ratings from the comprehension questions participants answered after each trial. The goal of this analysis is to ensure that participants were paying attention to the task, and so I will only be reporting results from comprehension questions following experimental trials, not fillers.

Accuracy ratings were high, with the overall percentage of correct responses being above 85% (see Figure 5.1). A linear mixed effect model using ADVERB TYPE and NUMBER SIZE as fixed effects, and SUBJECTS and ITEMS as random effects was fit to predict the accuracy ratings. Results show that there was no main effect of ADVERB TYPE (all p's > 0.1) or any significant ADVERB TYPE:NUMBER SIZE interactions (all p's > 0.2).

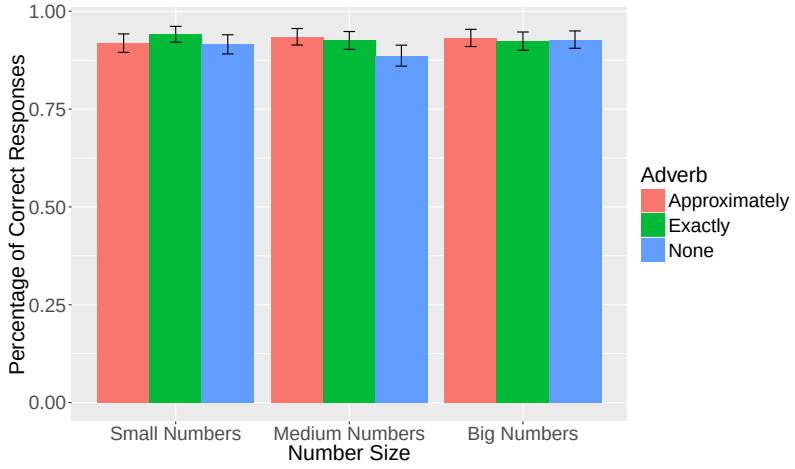


Figure 5.1: Accuracy ratings for the nine conditions tested in Experiment 3a.

Reading Times

Observations above 1000 ms and below 100 ms were removed from the data set submitted for statistical analyses (see figures 5.2 and 5.3 for the plots showing the raw Reading Times).

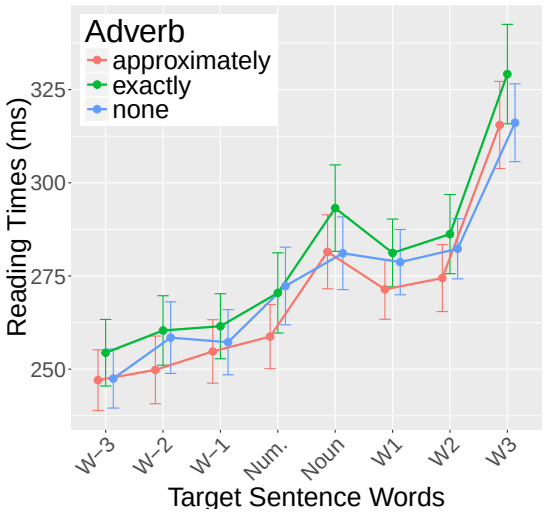


Figure 5.2: General Results for Experiment 3a. The vertical axis plots the raw Reading Times pertaining to the target sentences of all experimental trials collapsed over Number Size. The horizontal axis plots the Number and Noun regions, as well as three pre- and post-critical words.

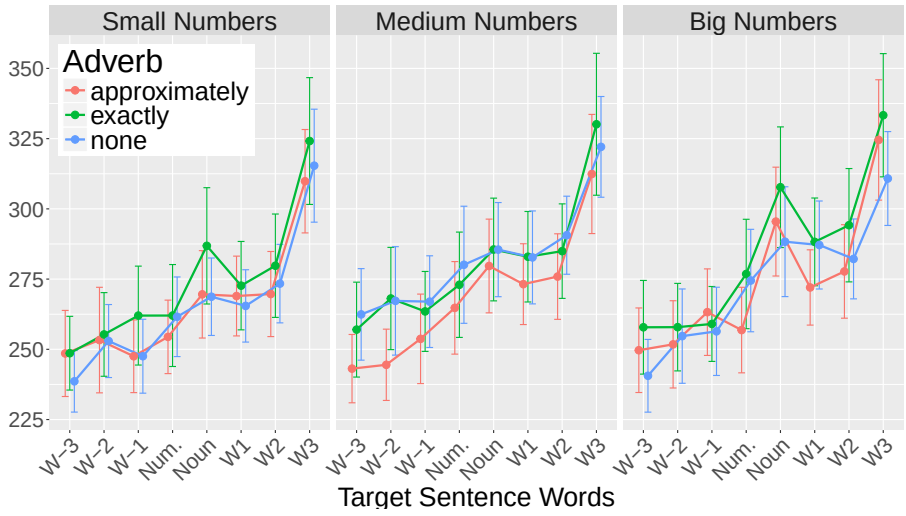


Figure 5.3: Raw Reading Times for Experiment 3a broken down by Number Size. The vertical axis plots the raw Reading Times pertaining to the target sentences of all experimental trials. The horizontal axis plots the Number and Noun regions, as well as three pre- and post-critical words.

Analyses were performed on the Log converted residual Reading Times obtained by running a mixed effects model to predict raw Reading Times. The model used as fixed effects 1) the word length of each of the words in the target sentence, and 2) the word's position in the sentence; and SUBJECTS as a random effect. The residuals of this model were later transformed to a logarithmic scale. All the Reading Time analyses reported in this chapter made used of this new dependent measure.

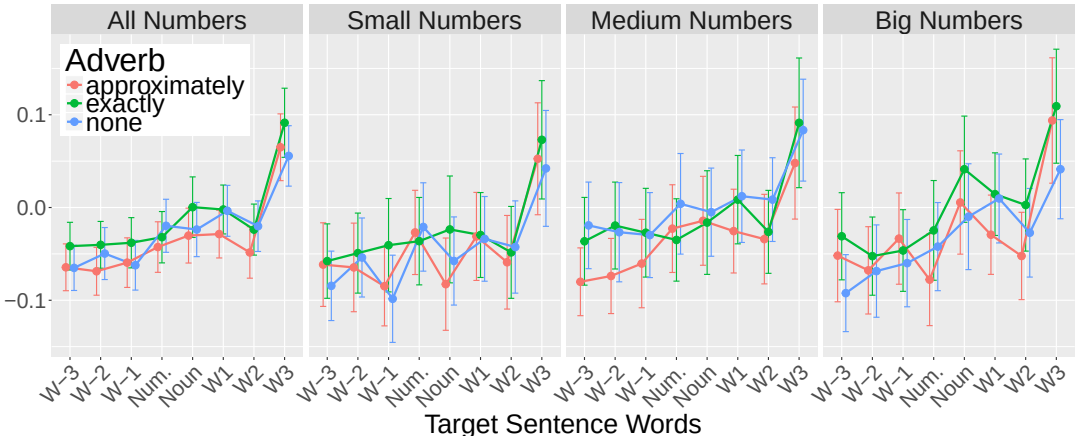


Figure 5.4: Log Residual Reading Times for the Number and Noun regions of the the target sentences pertaining to all experimental trials, as well as three pre- and post-critical words of Experiment 3a.

A linear mixed effects model was constructed using ADVERB TYPE and NUMBER SIZE as main effects. Furthermore, spillover effects on each of the target sentence words from two previous regions (SPILLOVER$_{w_{-1}}$ and SPILLOVER$_{w_{-2}}$) were also included in the model.[1] Finally, SUBJECTS and ITEMS were included as random effects. The model was run over the subset of the data pertaining to the Number and the Noun regions. Results did not reveal any significant effects of ADVERB TYPE (all p's > 0.4), NUMBER SIZE (all p's > 0.1), or any significant interactions (all p's > 0.7).[2]

1. The first and the second words of the target sentence were preceded by no words or just one word respectively. Therefore, the first word in the sentence did not have a value for either of these two predictors, whereas the second word in the sentence only had a value for predictor SPILLOVER$_{w_{-1}}$.

2. As expected, the two spillover predictors were highly significant. In the remainder of this chapter, I will not

5.2.4 Discussion

Experiment 3a failed to detect any differences across conditions as a function of ADVERB TYPE in either of the two critical regions, i.e. the Numeral and the Number region. Given that the accuracy ratings were consistently high, this lack of an effect cannot be attributed to participants not paying attention to the task. One possibility is that participants did not notice the slack regulator in the context, in which case no differences between conditions are expected to arise, since the context paragraphs are identical across conditions with the exception of the slack regulator. In order to control for this possibility, Experiment 3b tests the same contexts and target sentences used in Experiment 3a with the only difference that the slack regulator appears in the target sentence immediately preceding the numeral, not in the context paragraph.

5.3 Experiment 3b

5.3.1 Methods

Design and Materials

As in Experiment 3a, Experiment 3b employs a fully crossed 3(ADVERB TYPE: *approximately* vs. *Exactly* vs. *None*) x 3(NUMBER SIZE: *Small* vs. *Medium* vs. *Big*) design resulting in the nine conditions illustrated in (67). The crucial difference between Experiments 3a and 3b is that the adverb appears in the target sentence, as shown in (67b).

(67) a. **Context:** When the doctor arrived that morning she asked the nurse how many patients were waitlisted for the new clinical study.

b. **Target Sentence:** The nurse replied that {**approximately/ exactly/ ∅**} {**10/ 100/ 1,000**} **patients** were currently waitlisted.

report results belonging to these two predictors, even though they will be consistently included in all models.

c. **Comprehension Question:** When did the conversation between the nurse and the doctor take place?

d. **Possible Answers:** 1) Morning; 2) Night; 3) Afternoon.

The same 33 fillers used in Experiment 3a were included in Experiment 3b.

Apparatus

Same as Experiment 3a.

Procedure

Same as Experiment 3a.

Participants

Participants were 78 native speakers of American English between the ages of 18 and 35 years old (42 females, mean 21.8, range 18-35). All participants were recruited at the University of Chicago and were payed $10 for participating in the study. All subjects were naïve to the purpose of the experiment.

5.3.2 Predictions

Same as Experiment 3a.

5.3.3 Results

Accuracy Ratings

As was the case for Experiment 3a, accuracy ratings were consistently high for all the experimental conditions, with the overall percentage of correct responses being above 90% (see Figure 5.5). A

threshold of 70% of correct responses was set as the criterion to detect and subsequently discard data from subjects that were not following the instructions. Exploration of the accuracy ratings by subject revealed that only one participant achieved less than 70% of correct responses. Data belonging to this participant was excluded from data analysis. The same analysis procedure followed to analyze the accuracy ratings of Experiment 3a was applied to the accuracy data of Experiment 3b. As expected given the results reported for Experiment 3a, there were no significant main effects of ADVERB TYPE (all p's > 0.3), or any significant ADVERB TYPE : NUMBER SIZE interactions (all p's > 0.1), indicating that participants were paying attention to the task.

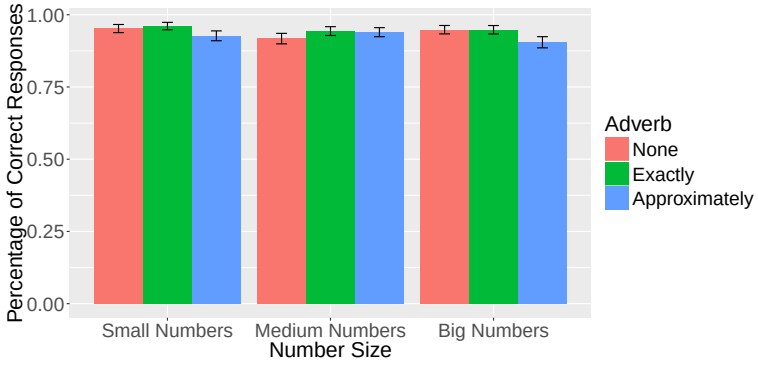


Figure 5.5: Accuracy Ratings for the nine experimental conditions tested in Experiment 3b.

Reading Times

Analyses were performed on a trimmed subset of the original data set after removing observations above 1,000 ms and below 100 ms. Data from the participant with low accuracy responses was also removed from the analyzed data set. Figures 5.6 and 5.7 show the raw Reading Times for the Adverb and the two critical regions (i.e. Numeral and Noun regions), as well as three pre- and post-critical regions, after trimming the data set.

The same mixed effects model described in §5.2.3 was used to predict the Log residual Reading Times (see Figures 5.8 and 5.9) of the Numeral and the Noun regions. I first focus on the analysis of the Numeral Region. Full statements of the results are given in Table 5.1, where I present the Type 3 analysis of variance (ANOVA) table for the fixed effects included in the model, with degrees

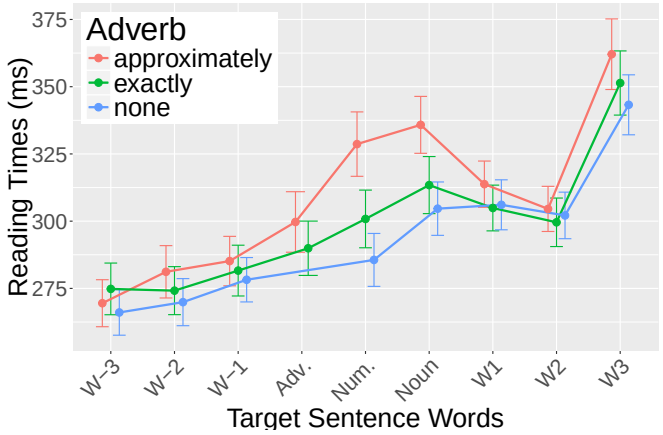


Figure 5.6: General Results for Experiment 3b. The vertical axis plots the raw Reading Times pertaining to the target sentences of all experimental trials collapsed over Number Size. The horizontal axis plots the Adverb, Number and Noun regions, as well as three pre- and post-critical words.

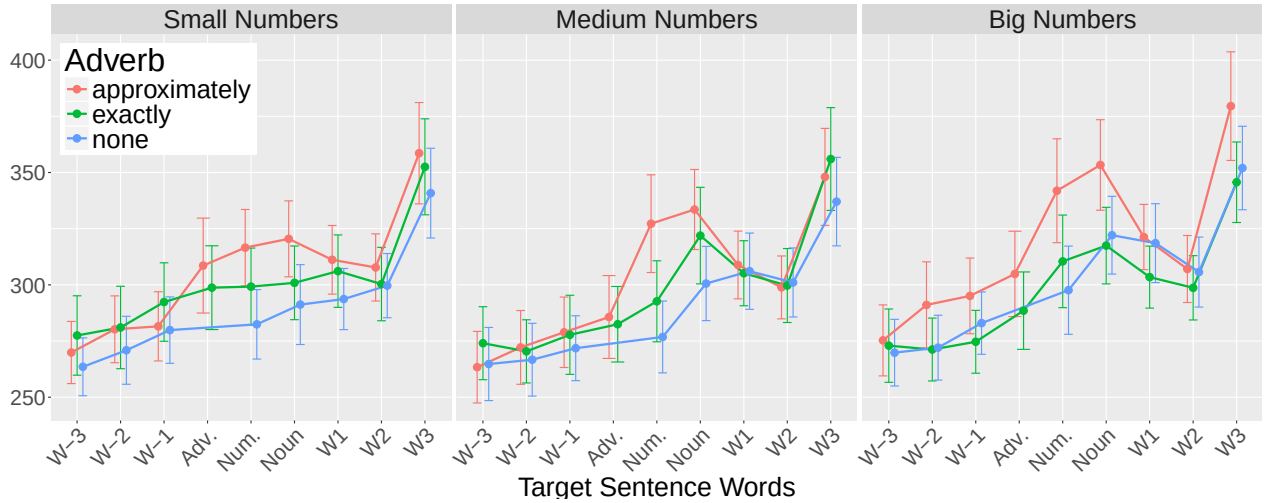


Figure 5.7: Raw Reading Times for Experiment 3b broken down by Number Size. The vertical axis plots the raw Reading Times pertaining to the target sentences of all experimental trials. The horizontal axis plots the Adverb, Number and Noun regions, as well as three pre- and post-critical words.

of freedom, F-value, and corresponding p-value calculated using Satterthwaite's approximation (using the lmerTest package in R). Table 5.2 summarizes the fixed-effect coefficients of the predictors included in the model. Coefficient significances were computed using t-tests with degrees of freedom calculated using the Satterthwaite approximation (again using lmerTest package).

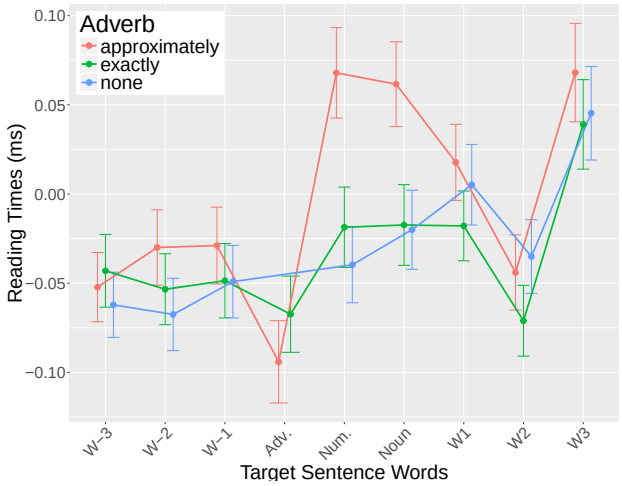


Figure 5.8: Log Residual Reading Times collapsed over NUMBER SIZE for Experiment 3b.

Figure 5.9: Log Residual Reading Times broken down by NUMBER SIZE for Experiment 3b.

Results revealed a main effect of ADVERB TYPE ($F(2, 46.08) = 10.736$, $p < 0.001$), such that the Reading Times (RTs) in the *approximately*-condition were significantly longer than the RTs in the No-Adverb condition for all three number sizes (SMALL NUMBERS: $\beta = 0.0731$, $SE = 0.0233$, $t = 3.132$, $p < 0.01$; MEDIUM NUMBERS: $\beta = 0.1215$, $SE = 0.0262$, $t = 4.629$, $p < 0.001$; BIG NUMBERS:

$\beta = 0.0593$, $SE = 0.0269$, t = 2.205, $p < 0.05$). The comparison between the *exactly* and the *approximately*-conditions also yield significant results, again due to the fact that participants spent more time reading the numeral in the latter condition compared to the former (SMALL NUMBERS: $\beta = 0.0645$, $SE = 0.0232$, t = 2.778, $p < 0.01$; MEDIUM NUMBERS: $\beta = 0.0812$, $SE = 0.0264$, t = 3.074, $p < 0.01$; BIG NUMBERS: $\beta = 0.0502$, $SE = 0.0267$, t = 1.878, $p = 0.06$). Finally, the No-Adverb condition vs. the *exactly*-condition comparison did not reach significance in any of the three number sizes tested (all p's > 0.1). No significant interactions emerged in this region ($p > 0.5$).

Predictor	Sum Sq	Mean Sq	NumDF	DenDF	F	p
NUMBER SIZE	0.0271	0.0136	2	26.98	0.310	> 0.7
ADVERB TYPE	0.9409	0.4705	2	46.08	10.736	< **0.001**
SPILLOVER$_{w_{-1}}$	6.7535	6.7535	1	1283.20	154.122	< 0.001
SPILLOVER$_{w_{-2}}$	6.7758	6.7758	1	1287.66	154.632	< 0.001
NUMBER SIZE : ADVERB TYPE	0.1451	0.0363	4	1068.62	0.828	> 0.5

Table 5.1: Analysis of Variance table for the Log Residual Reading Times pertaining to the Numeral Region of Experiment 3b, with F statistic, degrees of freedom, and p-values calculated using Satterthwaite's approximation.

To determine whether there were any effects of imprecision on the Residual RTs in the Noun region, the same model used to analyze the data belonging to the Numeral region was fit to the data of the Number region. Neither the main effect of ADVERB TYPE nor the ADVERB TYPE : NUMBER SIZE interaction were significant (all p's > 0.1). In order to investigate whether any of the effects found in the Numeral region could be attributed to spillover effects originating in processing differences between the two slack regulators tested, the data pertaining to the Adverb region was analyzed using a model that included ADVERB TYPE, NUMBER SIZE and their interactions as fixed effects, as well as random effects for SUBJECTS and ITEMS. No significant difference between the two adverbs was detected ($\beta = 0.02$, $SE = 0.03$, $p > 0.3$).

LogRTRes ~ NumberSize * AdverbType + Spillover$_1$ + Spillover$_2$ + (1 + NumberSize + AdverbType | Subject) + (1 + NumberSize + AdverbType | Item)				
Predictor	Coef β	SE(β)	t	p
Intercept	−3.631	0.1487	−24.421	< 0.001
Medium Numbers	−0.02789	−0.02610	−1.068	> 0.2
Big Numbers	0.01124	-0.02555	0.440	> 0.6
Exactly	−0.004694	0.02567	−0.183	> 0.8
Approximately	0.0604	0.02537	2.383	< **0.01**
Spillover$_{w_{-1}}$	0.3173	0.02556	12.415	< 0.001
Spillover$_{w_{-2}}$	0.307	0.02659	12.435	< 0.001
Medium Numbers : Exactly	0.04096	0.03419	1.198	> 0.2
Big Numbers : Exactly	0.0190	0.03415	0.319	> 0.7
Medium Numbers : Approximately	0.04474	0.03405	1.314	> 0.1
Big Numbers : Approximately	−0.008893	0.03419	−0.260	> 0.7

Table 5.2: Summary of fixed effects for the mixed effects linear regression model predicting Log Residual Reading Times belonging to the Number Region of Experiment 3b. The table contains coefficient estimates (β), standard errors, associated t-statistics, and significances. Significances below the *alpha* = 0.05 level for the relevant predictors, with the exception of the two Spillover predictors, are bolded.

5.3.4 Discussion

Results showed a clear penalty for imprecise interpretations of round numerals in the Number region, as the RTs for the numeral were significantly higher in the *approximately*-condition, compared to both the *exactly* and the *no-adverb* conditions. The results also show that the main effect of Adverb Type was uniquely driven by the extra processing cost associated with the imprecise interpretation of the numeral, since no significant differences were detected between the *exactly* and the *no-adverb* conditions. The fact that these two conditions patterned indistinguishably also suggests that the bare numeral was interpreted precisely.

The absence of a processing advantage for imprecise interpretations of round numbers indicates that interpreters were not biased towards approximate interpretations. Furthermore, the *no-adverb* condition patterned like the *exactly*-condition and unlike the *approximately*-condition, suggesting

that bare numerals were interpreted precisely. This fact is difficult to reconcile with the *Coarsest Scale Principle*, which poses a preference for approximation whenever possible.[3] The present results did not only fail to detect a processing advantage for imprecision, they also uncovered a processing penalty for imprecise interpretations of number words compared to precise ones. Moreover, the fact that the *no-adverb* condition behaved like the *exactly*-condition and unlike the *approximately*-condition also suggests that the default interpretation of bare numerals is precise. Taken together, these findings suggest that when it comes to the interpretation of round numerals precision is generally favored and easier to process when compared to imprecise interpretations. Finally, the absence of an interaction between approximation and the magnitude of the numeral goes against Krifka's (2007) claim that bigger numbers should lend themselves to approximate interpretations more easily because they are represented in coarser-grained granularity scales than smaller numbers. The absence of a significant interaction is perhaps to be expected, since in Krifka's account, the asymmetry between small and bigger numbers follows from the preference for the coarsest-grained interpretation licensed in the context, a preference that does not seem to be borne out by the current results.

3. The latter conclusion must at this point be taken with caution. One possibility is that the majority of the tested stimuli did not license imprecise interpretations of the numerals. In this case, the current results would not constitute evidence against the claim that there exists a preference for approximation, as stated by the *Coarsest Scale Principle*, since such preference only applies to discourses where imprecise interpretations are available to begin with. I further discuss this point in the context of Experiment 5. A second possibility is that bare numerals were in fact interpreted imprecisely, while at the same time, there was a processing advantage for such interpretation, resulting in the *no-adverb* condition patterning very similarly to the *exactly*-condition. If this was the case, the fact that the RTs in the Number region were comparable in the *exactly* and the *no-adverb* condition would not indicate that the bare numerals were interpreted precisely. This pattern of results would suggest that the processing of imprecision is fundamentally different depending on the nature of the cues, i.e. semantic vs. pragmatic, comprehenders make use of in order to arrive at and construct an imprecise interpretation. Unfortunately, these questions cannot be answered through the current experiment, since no judgement data about the actual interpretation of the bare numerals was collected. In Experiment 6b (§5.7), presented below, seeks to tease apart these two possibilities.

The current experiment contained the slack regulator in the target sentence immediately preceding the numeral (recall that experiment 3a, which tested the same target sentences with the slack regulators appearing in the context paragraph did not yield any significant differences among conditions). Given the close proximity between the slack regulator and the numeral, an important question is how much of the extra processing cost observed in the *approximately*-condition are spillover effects stemming from the processing of the adverb itself. Even though no significant differences arose between the *exactly* and the *approximately*-conditions in the adverb region, it is conceivable that factors like the lexical frequency of these two adverbs might have affected the processing of subsequent regions. In order to address this concern, I used the Corpus of Contemporary American English (COCA), which contains more than 520 million words, to obtain the probability of each of the slack regulators tested given the occurrence of a number word. The frequency of the bigram [*approximately*+numeral] was < 0.5%, whereas the bigram frequency of [*exactly*+numeral] was < 0.03%.[4] While the relative frequency of *approximately* given a numeral is higher than the relative frequency of *exactly*, the absolute frequency of *approximately* is much lower, i.e. 19,761 instances found in COCA, than that of *exactly*, i.e. 72,130 instances. Therefore, it is conceivable that the lower absolute frequency of the adverb *approximately* could be incurring higher processing costs and therefore partially or fully driving the effects reported in §5.3.3. In Experiment 4 I address this issue by substituting the slack regulator *approximately* with *about*, an adverb whose relative and absolute frequency are much higher than that of *exactly*.

5.4 Experiment 4

The goal of Experiment 4 is to determine to what extent the effects detected in Experiment 3b were the result of extra processing demands linked to the lower absolute frequency of the slack regulator *approximately* compared to the slack regulator *exactly*. Experiment 4 replicates the design

4. These calculations only take into consideration arabic representations of number words, not alphabetical representations. Only numbers within the range tested in the experiment, i.e. 100 to 90,000, where included in this search.

of Experiment 3b substituting the adverb *approximately* by the adverb *about*, which is attested 144,4147 times in the COCA corpus, a much higher frequency than that of *exactly*. This count does not discriminate between uses of *about* as an adverb (e.g. 'It was *about* 3 o'clock.') from uses of *about* as a preposition (e.g. 'It's all *about* publishing.'). The frequency of *about* when relativized to a numeral is of < 4%. Therefore, when compared to *exactly*, which has an absolute lexical frequency of 72,130 and a bigram frequency of < 0.03%, *about* is more frequent by both measures. If the effects observed in Experiment 3b did not reflect the processing of imprecision but rather the difficulty of processing a lower frequency adverb, we should not find any differences in Experiment 4 between the *about*- and the *no-adverb*/*exactly*-conditions.

5.4.1 Methods

Design and Materials

The design and materials are the same as those used in Experiment 3b with the exception that the Slack Regulator *approximately* was substituted by the slack regulator *about*. An item example can be found in (68).

(68) a. **Context:** When the doctor arrived that morning she asked the nurse how many patients were waitlisted for the new clinical study.

b. **Target Sentence:** The nurse replied that {**about/ exactly/ ∅**} {**10/ 100/ 1,000**} **patients** were currently waitlisted.

c. **Comprehension Question:** When did the conversation between the nurse and the doctor take place?

d. **Possible Answers:** 1) Morning; 2) Night; 3) Afternoon.

Apparatus and Procedure

Same as Experiments 3a/b.

Participants

Participants were 68 native speakers of American English between the ages of 18 and 35 years old (42 females, mean age 19, range 18-24). All participants were recruited at the University of Chicago and were either payed $10 for participating in the study or were granted research credits. All participants were naïve to the purpose of the experiment. No participants were excluded from data analysis.

5.4.2 Predictions

If the results of Experiment 3b were due to an extra processing load resulting from the processing of approximate interpretations of round numbers, we expect to find a similar pattern of effects in Experiment 4. On the other hand, to the extent that the increased RTs observed in the Numeral region for the *approximately*-condition of Experiment 3b were due to the lower frequency of the adverb *approximately*, no extra processing cost should be detected in the *about*-condition of the current experiment, since the slack regulator *about* has a much higher frequency than the slack regulator *approximately*.

5.4.3 Results

Accuracy Ratings

As in the previous experiments, accuracy ratings were very high, with all conditions receiving mean accuracy ratings above 90% (see Figure 5.10). A mixed effect model using ADVERB TYPE and NUMBER SIZE and its interaction as predictors, and SUBJECTS and ITEMS as random effects, confirmed that there were no significant differences in the ratings of the nine conditions tested in the experiment (all p's < 0.1).

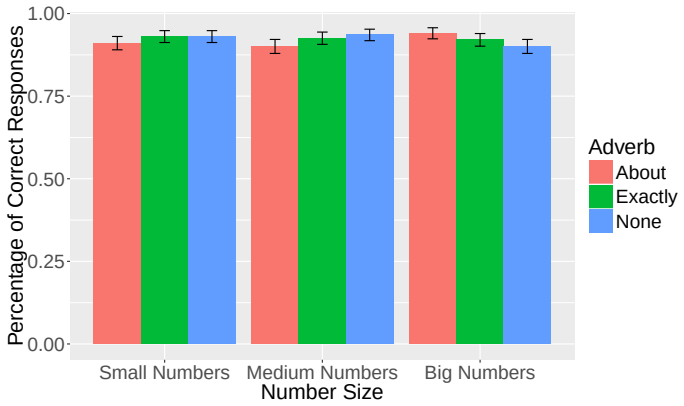


Figure 5.10: Accuracy Ratings for the nine Experimental Conditions tested in Experiment 4.

Reading Times

Raw RTs for the nine conditions included in Experiment 4 are plotted in Figure 5.11. The same models constructed for the statistical analysis of the results of Experiment 3b were used to predict the Log residual RTs (see Figure 5.12) of the Adverb, Number and Noun regions.

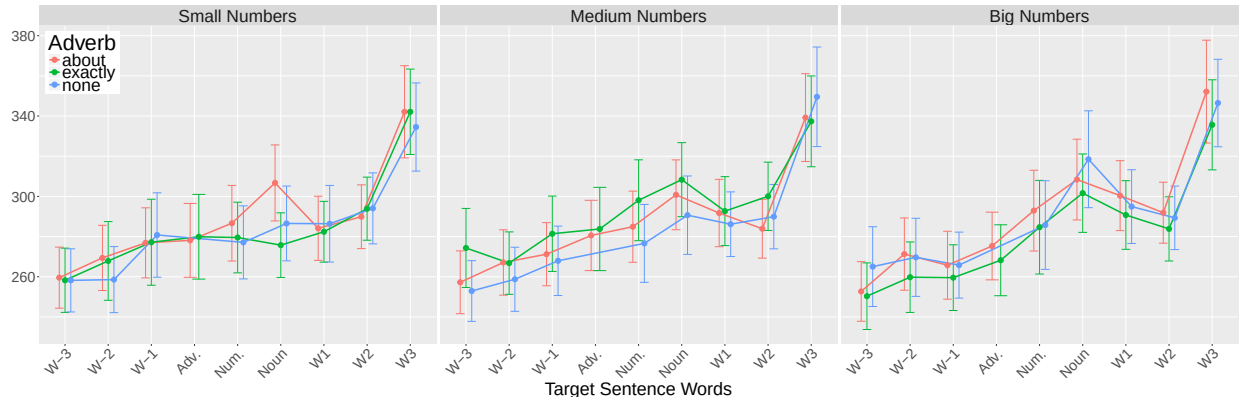


Figure 5.11: Raw Reading Times for Experiment 4 broken down by Number Size. The horizontal axis plots the three regions of interest, i.e. the Adverb, Numeral and Noun regions, plus the three words preceding and following the three regions of interest.

In the Adverb region, results showed no significant difference between the *exactly* and the *about*-conditions ($\beta = -0.01363$, $SE = 0.02539$, $p > 0.593$). Similarly, in the Numeral region no significant main effect of ADVERB TYPE, or ADVERB TYPE : NUMBER SIZE interaction was detected

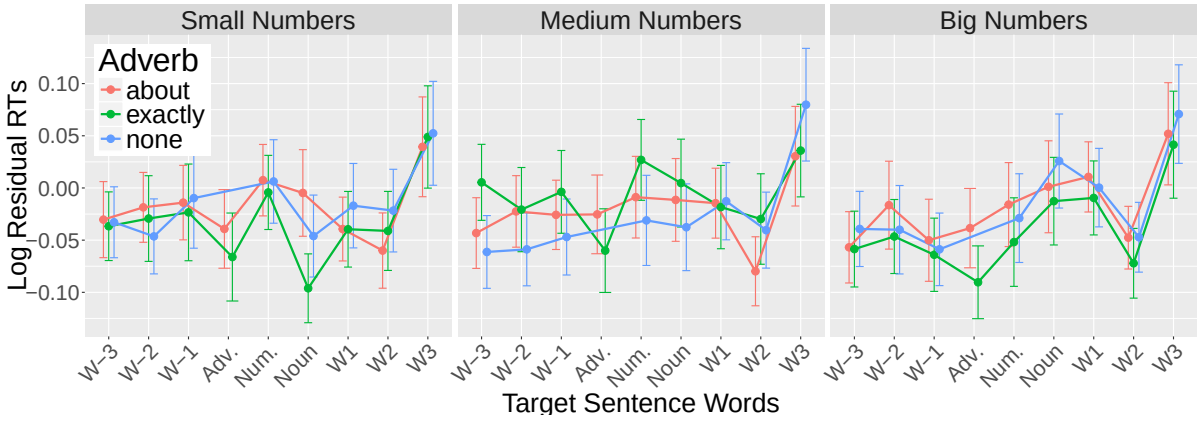


Figure 5.12: Log Residual Reading Times for Experiment 4. The horizontal axis plots the three regions of interest, i.e. the Adverb, Numeral and Noun regions, plus the three words preceding and following these critical regions of interest.

(all p's > 0.2).

For the Noun region, analyses did not reveal a main effect of ADVERB TYPE ($F(2, 36.92) = 0.879$, $p > 0.4$). However, there was a main effect of NUMBER SIZE ($F(2, 117.21) = 5.709$, $p < 0.005$) and a significant ADVERB TYPE : NUMBER SIZE interaction ($F(4, 1072.64) = 2.737$, $p < 0.03$, see Figure 5.3). Further comparisons showed that the main effect of NUMBER SIZE was driven by the difference between small and big numbers ($\beta = 0.0777$, $SE = 0.0259$, $p < 0.003$), not by the difference between small and medium numbers ($\beta = 0.0234$, $SE = 0.0254$, $p > 0.3$).

Predictor	Sum Sq	Mean Sq	NumDF	DenDF	F	p
NUMBER SIZE	0.53116	0.26558	2	117.21	5.709	**0.004307**
ADVERB TYPE	0.08175	0.04088	2	36.92	0.879	0.423855
SPILLOVER$_{w_{-1}}$	2.98786	2.98786	1	1191.22	64.225	< 0.00001
SPILLOVER$_{w_{-2}}$	2.49247	2.49247	1	1270.84	53.577	< 0.00001
NUMBER SIZE : ADVERB TYPE	0.50933	0.12733	4	1072.64	2.737	**0.027679**

Table 5.3: Analysis of Variance table for Log Residual Reading Times pertaining to the Noun Region of Experiment 4. The table shows the F-statistic, degrees of freedom, and p-values calculated using Satterthwaite's approximation.

To explore the ADVERB TYPE : NUMBER SIZE interaction, I examined the effect of ADVERB

TYPE on the three number sizes tested separately, using both the *no-adverb* condition (Tables 5.4-5.6) and the *exactly*-condition as the baseline for comparison (Tables 5.7-5.9). Results showed a processing advantage for the *exactly*-condition compared to the *about*-condition for Small Numbers (β = 0.08577, *SE* = 0.025, t = 3.473, $p < 0.001$, Table 5.7), whereas the same comparison was not significant neither for Medium nor for Big Numbers (all p's > 0.2). The comparison between the *exactly*-condition and the *no-adverb* condition, showed a trend in the same direction for Small Numbers, such that the RTs were faster in the *exactly*-condition than in the *no-adverb* condition (β = 0.04584, *SE* = 0.025, t = 1.862, $p < 0.07$, Table 5.7). Again, this trend fades for bigger numbers (all p's > 0.3). Finally, the comparison between the *about*-condition and the *no-adverb* condition did not reach significance in either of the three number sizes (all p's > 0.1).

LogRTRes ~ ADVERBTYPE + (1 + ADVERBTYPE | Subject) + (1 + ADVERBTYPE | Item)

Predictor	Coef β	SE(β)	**t**	p
Intercept	−1.42666	0.17155	−8.316	< 0.0001
Exactly	−0.04556	0.02481	−1.836	0.0689
About	0.04027	0.02569	1.567	0.1221
SPILLOVER$_{w_{-1}}$	0.07379	0.04399	1.677	0.0943
SPILLOVER$_{w_{-2}}$	0.17439	0.03969	4.394	< 0.0001

Table 5.4: Full mixed effects linear regression model coefficients for the Small Numbers of the Noun Region of Experiment 4, using the *no-adverb* condition as baseline for comparison.

LogRTRes ~ ADVERBTYPE + (1 + ADVERBTYPE | Subject) + (1 | Item)

Predictor	Coef β	SE(β)	**t**	p
Intercept	−0.979109	0.197821	−4.949	< 0.0001
Exactly	0.026763	0.028083	0.953	0.341274
About	0.019236	0.027963	0.688	0.492046
SPILLOVER$_{w_{-1}}$	0.166419	0.042959	3.874	< 0.0001
SPILLOVER$_{w_{-2}}$	0.003368	0.043548	0.077	0.938394

Table 5.5: Full mixed effects linear regression model coefficients for the Medium Numbers of the Noun Region of Experiment 4, using the *no-adverb* condition as baseline for comparison.

LogRTRes ~ AdverbType + (1 + AdverbType | Subject) + (1 + AdverbType | Item)				
Predictor	Coef β	SE(β)	**t**	p
Intercept	−1.46494	0.21123	−6.935	< 0.0001
Exactly	−0.03954	0.04027	−0.982	0.33776
About	−0.03708	0.03141	−1.180	0.25512
SPILLOVER$_{w_{-1}}$	0.09588	0.04275	2.243	0.02549
SPILLOVER$_{w_{-2}}$	0.17340	0.05400	3.211	< 0.002

Table 5.6: Full mixed effects linear regression model coefficients for the Big Numbers of the Noun Region of Experiment 4, using the *no-adverb* condition as baseline for comparison.

LogRTRes ~ AdverbType + (1 + AdverbType | Subject) + (1 | Item)				
Predictor	Coef β	SE(β)	**t**	p
Intercept	−1.47811	0.17178	−8.605	< 0.0001
None	0.04584	0.02462	1.862	0.064418
About	0.08577	0.02469	3.473	**0.000627**
SPILLOVER$_{w_{-1}}$	0.07611	0.04406	1.728	0.084887
SPILLOVER$_{w_{-2}}$	0.17312	0.03971	4.360	< 0.0001

Table 5.7: Full mixed effects linear regression model coefficients for the Small Numbers of the Noun Region of Experiment 4, using the *exactly*-condition as baseline for comparison.

LogRTRes ~ NumberSize * AdverbType + (1 + AdverbType | Subject) + (1 + AdverbType | Item)				
Predictor	Coef β	SE(β)	**t**	p
Intercept	−0.936369	0.199092	−4.703	< 0.0001
None	−0.026094	0.034130	−0.765	0.453
About	−0.006159	0.037742	−0.163	0.872
SPILLOVER$_{w_{-1}}$	0.170402	0.042506	4.009	< 0.0001
SPILLOVER$_{w_{-2}}$	−0.003680	0.043144	−0.085	0.932

Table 5.8: Full mixed effects linear regression model coefficients for the Medium Numbers of the Noun Region of Experiment 4, using the *exactly*-condition as baseline for comparison.

LogRTRes ~ NUMBERSIZE * ADVERBTYPE + (1 + ADVERBTYPE | Subject) + (1 + ADVERBTYPE | Item)				
Predictor	Coef β	SE(β)	**t**	p
Intercept	−1.46494	0.21123	−6.935	< 0.0001
None	−0.03954	0.04027	−0.982	0.33776
About	−0.03708	0.03141	−1.180	0.25512
SPILLOVER$_{w_{-1}}$	0.09588	0.04275	2.243	0.02549
SPILLOVER$_{w_{-2}}$	0.17340	0.05400	3.211	0.00144

Table 5.9: Full mixed effects linear regression model coefficients for the Big Numbers of the Noun Region of Experiment 4, using the *exactly*-condition as baseline for comparison.

5.4.4 Discussion

Consistent with the findings of Experiment 3b, Experiment 4 also uncovered an asymmetry between the processing of precise and imprecise interpretations of round numbers. However, there exist some important differences between the patterns of the effects found in Experiment 3b and the current experiment. First, in Experiment 4 the effect presented as an interaction between the factors ADVERB TYPE and NUMBER SIZE, whereas in Experiment 3b the penalty of imprecision emerged in the form of a main effect. Second, the interaction did nook take place in the Numeral region, where the effects for Experiment 3b were detected, but rather in the Noun region. Experiments 4 did however converge with Experiment 3b in that no significant differences between the two slack regulators arose in the Adverb region.

The interaction found in Experiment 4 was driven by the significant difference between the *about*-condition and the *exactly*-condition, and the marginal difference between the *no-adverb* condition and the *exactly*-condition, both of which were found for Small Numbers, but not for Medium and Big Numbers. Coupled with the lack of an effect for the *no-adverb* vs. *about*-condition comparison in the same region, these results suggest that the bare numerals in the *no-adverb* condition received a more precise interpretation in Experiment 4 than in Experiment 3b. Unfortunately, this is a question that cannot be addressed, at least for the current experiment, since none of the data

collected speaks to the question of the degree of precision with which participants interpreted the bare round numeral.[5] In any case, based on the results of the Noun Region, the effects of Experiment 4 cannot be construed as a penalty for imprecise interpretations alone, as was the case for Experiment 3b. Rather, it seems that the effect was driven by the combination of a processing penalty for imprecise interpretations (albeit not significant if the *no-adverb* condition is taken as the benchmark for comparison), and a processing advantage for precise readings of round numbers.

Importantly, the differences between Experiments 3b and 4 suggest that the lower absolute frequency of the adverb *approximately* compared to that of the adverb *exactly* indeed had an effect over the RTs of the imprecise condition of Experiment 3b, probably causing the slow down of the RTs in the *approximately*-condition to generalize to all number sizes. It is also possible that the lower frequency of the adverb *approximately* caused the effect to be more immediate and appear in the Numeral region, as opposed to the Noun region as was the case for Experiment 4.

Abstracting away from these differences, results from Experiment 4 show that even after testing the slack regulator *about*, which has a much higher absolute frequency than *exactly*, there still persists a processing asymmetry between imprecise interpretations of round numbers (tested through the *about*-condition) and their precise counterparts (tested through the *exactly*-condition), such that the latter were faster to process than the former. These findings are difficult to reconcile with the view that imprecision is cognitively less costly than precision, at least from the addressee's perspective. The current results point at precisely the opposite direction, namely that precise interpretations are easier to process, at least when explicitly signaled with a slack regulator.

Finally, the fact that in Experiment 4 the effect emerged in the form of an interaction deserves some more elaboration. Recall that in Experiment 3b, despite the existence of a main effect of ADVERB TYPE, the penalty for imprecise readings was weaker for Big Numbers than it was for Small and Medium Numbers, at least when imprecise interpretations were compared to precise

5. Though see the experimental design of Experiments 6a and 6b below, where this issue was resolved by collecting both RTs and judgements about the perceived precision level of the numeral for each participant and experimental trial.

ones. Experiment 4 results, again, suggest that imprecise interpretations are faster to process for Big Numbers than they are for Small Numbers, since no differences were found for Medium and Big numbers in the *exactly* vs. *about*-condition comparison. As discussed in the context of Experiment 3b, the *Coarsest Scale Principle* predicts that bigger numbers should more readily be assigned an imprecise interpretation as a direct consequence of an alleged preference for approximation. However, neither Experiment 3b nor Experiment 4 have provided any evidence for the existence of such interpretational bias, suggesting that any processing advantage for imprecise interpretations of bigger numbers, and/or any processing disadvantage for precise interpretations of bigger numbers cannot be the result of a bias towards imprecision.

5.5 Experiment 5

The *no-adverb* condition included in Experiments 3b and 4 was taken to represent the *default* interpretation for the numeral given the previous linguistic context. Experiments 3b and 4 did not address the question of what the interpretational preferences of the bare numeral were, given a) the preceding linguistic context, and b) the size of the numeral itself. For instance, it is conceivable that the precision level generally adopted by participants was not stable across the three number sizes tested. Similarly, certain items might bias participants towards a more precise interpretation of the numeral than others, a factor that at the same time could interact with other independent considerations that have been claimed to determine the likelihood of adopting an imprecise interpretation, e.g. number magnitude. To put an example, the critical regions of the *no-adverb* condition could reflect a precise interpretation of the number word for small numbers, and an approximate interpretation of the number word for larger numbers. Therefore, in order to accurately interpret the interaction found in Experiment 4 between approximation and the size of the numeral, it is important to assess whether any potential interpretational differences of the bare numerals were pronounced enough to affect the online RTs across number sizes. Experiment 5 presented below has the objective of determining the general precision-level adopted by the participants for the interpretation of the

experimental materials used as control in Experiments 3b and 4.

5.5.1 Methods

Design and Materials

Materials consisted of a subset of the conditions included in Experiments 3b and 4, namely the three conditions used as control, with a total of 27 experimental items. Therefore, the only factor tested in this experiment is NUMBER SIZE (Small vs. Medium vs. Big Numbers), resulting in the three conditions illustrated in (69).

(69) a. **Context:** When the doctor arrived that morning she asked the nurse how many patients were waitlisted for the new clinical study.

b. **Target Sentence:** The nurse replied that {10/ 100/ 1,000} patients were currently waitlisted.

c. **Question:** How likely do you think it is that the nurse meant that exactly {10/ 100/ 1,000} patients were currently waitlisted?

Thirty of the fillers included in Experiments 3b and 4 were also used in the current experiment.

Procedure

Participants received written instructions and performed three practice trials with the goal of ensuring that they understood the task. Next, they proceeded to the experiment. For each trial, participants read the context and corresponding target sentence as a single paragraph. After reading the paragraph, they continued to the next screen by pressing the space bar. Participants were then asked to rate on a 1-7 scale the likelihood of the precise interpretation of the numeral presented in the paragraph, with 1 being *very unlikely* and 7 being *very likely*. The question, exemplified in (69c), was always formulated in the same way, although it was specific to the particular trial.

Participants provided their answer by clicking on one of the numbers that formed the scale. The experiment lasted approximately 15 minutes.

Participants

Participants were 28 native speakers of English (15 female, mean = 29.97, range = 21-53) recruited through the online platform Mechanical Turk. Participants were payed $3.00 for their participation. No participants were excluded from data analysis.

5.5.2 Results

Figure 5.13 shows the mean precision ratings for the three conditions tested in Experiment 5. As can be seen in the plot, Small Numbers received the highest mean ratings indicating that they tended to be interpreted the most precisely of the three number sizes tested, whereas Big Numbers received the lowest ratings. A mixed effects model using NUMBER SIZE as a fix effect and SUBJECTS and ITEMS as random effects was fit to the data to predict ratings as a function of NUMBER SIZE. Results confirmed that there was an effect of NUMBER SIZE ($F(2, 46.58) = 8.5622$, $p < 0.001$), driven by a significant difference between the Small and the Medium Numbers ($\beta = -0.632$, $SE = 0.17$, $t = -3.65$, $p < 0.001$), as well as a difference between the Small and the Big Numbers ($\beta = -0.906$, $SE = 0.22$, $t = -3.96$, $p < 0.001$, see Table 5.10 for the full model). The only comparison that did not reach significance was the Medium vs. Big Numbers comparison ($\beta = 0.274$, $SE = 0.161$, $t = 1.69$, $p < 0.1$, see Table 5.11).

Ratings ~ NUMBERSIZE + (1 + NUMBERSIZE | Subject) + (1 + NUMBERSIZE | Item)				
	Coef β	SE(β)	t	p
Intercept	4.714	0.24	19.21	0.001
Medium Numbers	−0.632	0.17	−3.65	< **0.001**
Big Numbers	−0.906	0.22	−3.96	< **0.001**

Table 5.10: Full mixed effects linear regression model for Experiment 5 using Small Numbers as the baseline for comparison.

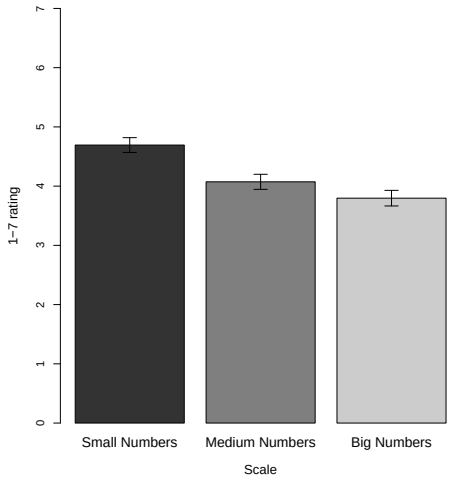


Figure 5.13: Mean precision ratings for the three control conditions used in Experiments 3b and 4.

Ratings ~ NUMBERSIZE + (1 + NUMBERSIZE | Subject) + (1 + NUMBERSIZE | Item)				
	Coef β	SE(β)	t	p
Intercept	3.808	0.305	12.4	< 0.001
Medium Numbers	0.274	0.161	1.69	< 0.1
Small Numbers	0.906	0.228	3.96	< **0.001**

Table 5.11: Full mixed effects linear regression model for Experiment 5 using Big Numbers as the baseline for comparison.

5.5.3 Post-hoc Analysis

In order to determine whether the differences in interpretation of the number words found in Experiment 5 were reflected in the RTs, I pulled the data belonging to the control condition of Experiment 4. For each number size, I divided the RTs in items that obtained a mean precision rating between 4 and 7 in Experiment 5, i.e. items that were rated as having a high precision bias (represented by the pink line in Figure 5.14), and items that received a low precision rating in Experiment 5, i.e. items that were rated between 1 and 4 (represented by the blue line in Figure 5.14).

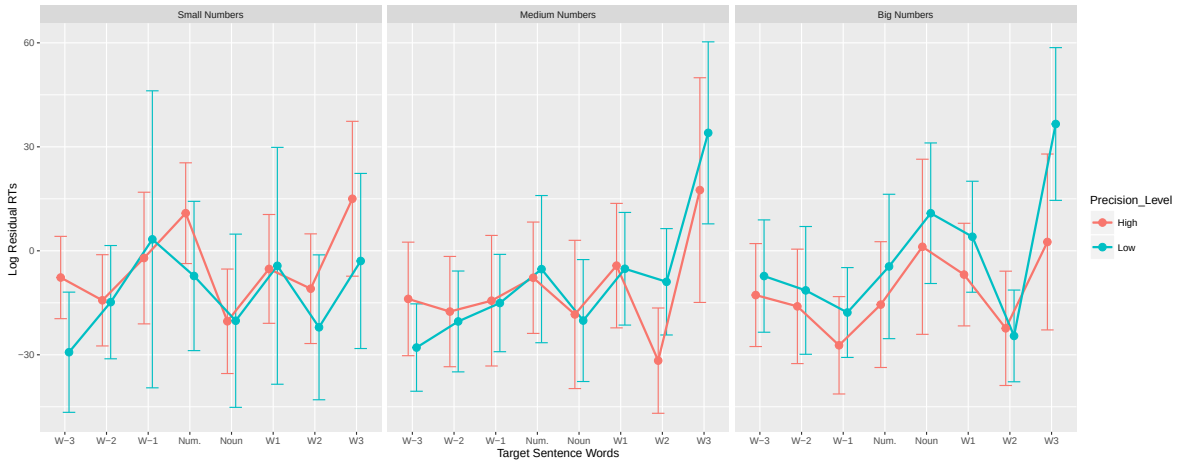


Figure 5.14: Reading Times belonging to the *no-adverb* condition of 4 for the three number sizes tested. The Reading Times are subdivided in items that were judged to have a high precision bias in Experiment 5 (labeled as 'High') versus items that were judged to have a low precision bias (labeled as 'Low').

Next, I constructed a model to determine whether there exist significant differences in the RTs of the Numeral and the Noun regions modulated by the precision level. Results did not show any significant differences in the Numeral Region (all p's > 0.1), and no significant effects in the Noun Region (all p's > 0.5).

5.5.4 Discussion

Results from Experiment 5 show that the perceived precision threshold was not stable across all three number sizes tested. Participants judged the likelihood that the speaker intended the round numeral to be interpreted imprecisely to be lower for Small numbers than for Medium and Big numbers. However, the differences found in the rating data were not paralleled by the RTs of the same condition, at least when the RTs were subdivided as described in §5.5.3. At this point, it is difficult to interpret the mismatch between the judgement and the RT data, but I will point out three possibilities. First, it is important to recall that the judgement and RT data come from two different pools of participants. If there exists a lot of individual variability in the interpretation of bare round

numerals, the mismatch between the ratings and the RTs could just reflect different interpretational preferences of the bare round number by the two groups of participants. Second, the interpretational differences across number sizes detected in the judgement data might not be pronounced enough to be mirrored in the RTs. After all, the numerical differences between the means of the three conditions are fairly small, see Figure 5.13. Finally, it is possible that the current results actually show that there is no relationship between the offline judgement data and the online RTs. This could be the case if the offline judgements are actually shaped by general pragmatic reasoning that is performed *after* the whole sentence is fully processed. In this case, the mismatch between the offline judgements and the RTs would be expected. In any case, the current results suggest that despite the fact that bigger bare numbers were perceived to be slightly more imprecise than smaller numbers, the RTs pertaining to the *no-adverb* condition of Experiment 4 were stable across number sizes, at least during the processing of the Numeral and the Noun regions. These findings enhance our interpretation of the interaction found in Experiment 4, as they confirm that the lack of significant differences for Medium and Big Numbers is not the result of fluctuations in the RTs of the *no-adverb* condition caused by considerable differences in the default interpretation of smaller vs. bigger bare numbers.

5.6 Experiment 6a

The goal of Experiment 6a is to determine whether the processing signature of imprecise meanings that come about as a result of pragmatic reasoning is the same as the processing of imprecise meanings that are induced by a slack regulator.

5.6.1 Methods

Participants

Participants were 30 native speakers of American English between the ages of 18-35 years old (12 female, mean=27.8, range=18-34) recruited at the University of Chicago. Participants were payed $10 for participating in the study. All participants were naïve to the purpose of the experiment.

Apparatus

Same as Experiments 3 and 4.

Procedure

Same as Experiments 3 and 4.

Design and Materials

The experiment followed a 2 x 3 factorial design. The first factor was the PRECISION BIAS, i.e. *High* vs. *Low*, used in the linguistic context preceding the numeral. The second factor, which was shared with Experiments 3-4, was NUMBER SIZE, and consisted of the same three levels, i.e. *Small* vs. *Medium* vs. *Big Numbers*, tested in Experiments 3-4.

A total of twenty-four experimental items were distributed in six lists using a Latin Square Design. Each item consisted of a linguistic context (70a-b) and a target sentence containing the numeral (70c). The target sentences were the ones used in Experiments 3-4. Three of the target sentences tested in Experiments 3-4 were not included in Experiment 6a because it proved difficult to construct contexts for them following the criteria described in this section. The last sentence used in the two precision bias contexts within one item was always identical to the linguistic contexts used in Experiments 3 and 4 (see example (70)). As in the previous self-paced reading experiments reported in this chapter, the Noun region following the numeral was always followed by at least

three more words in order to avoid sentence final wrap up effect confounds in the critical regions of interest, i.e. the numeral and the noun following the numeral. The same 33 fillers used in Experiments 3-4 were included in this experiment. All participants saw the same set of fillers.

(70) a. **High Precision Bias Context:** The doctor and the nurse were filling out the FDA forms for their most recent clinical trial. The doctor told the nurse that they needed to report the number of patients that ended up in the waitlist to the FDA, so she asked the nurse how many patients were currently waitlisted.

b. **Low Precision Bias Context:** While leaving the hospital on the way home, the doctor and the nurse were chatting about their latest clinical trial. The doctor told the nurse that she was thinking of piloting a new study with some of the patients in the waitlist, so she asked the nurse how many patients were currently waitlisted.

c. **Target:** The nurse replied that {10/ 100/ 1,000} patients were currently waitlisted.

The *high* and the *low* precision bias contexts manipulated elements like the goal that the speaker intended to achieve by asking the question answered by the target sentence, or the context in which the conversation was taking place. For instance, in example (70a) the doctor inquires about the number of waitlisted patients in order to report that information to the FDA, an official government agency that expects to receive both true and precise data. Therefore it should be reasonable for the nurse to assume that the doctor was expecting a precise answer to her question. On the other hand, such assumption is not likely to be in place in context (70b), where the doctor and the nurse are having a casual conversation while walking home. Therefore, the likelihood that the nurse intended his utterance of (70c) to be interpreted imprecisely should be much higher in the context of (70b) than in the context of (70a). In order to ensure that the pragmatic cues used in Experiment 6a led to the expected interpretation biases of the target numeral, all the experimental items were normed. In what follows, I provide the details of this norming study.

Norming Study

Participants. A total of 47 participants (16 females, mean=27.98, range=18-40) were recruited through the website Mechanical Turk. In exchange for their participation, participants were payed $3.00. All participants were naïve to the purpose of the study.

Materials and Procedure. Materials consisted of the 24 experimental items tested in Experiment 6a. Participants read the context and the target sentence as one single paragraph, and rated the likelihood that the speaker intended the numeral to be interpreted precisely by answering questions of the form of (71).

(71) How likely do you think it is that the nurse meant that exactly 10/100/1,000 patients were currently waitlisted?

Participants were provided a 1-7 scale, with 1 being *very unlikely* and 7 being *very likely* (this is the same exact task used in Experiment 5).

Results and Discussion. Results show that participants were sensitive to the contextual manipulations. As can be seen in Figure 5.15, the High Precision Bias condition systematically received higher ratings, indicating that participants interpreted the target numeral more precisely in the High Precision Bias condition than in the Low Precision Bias condition. A mixed effect model using the factors PRECISION BIAS and NUMBER SIZE as predictors of the ratings, and SUBJECTS and ITEMS as mixed effects confirmed that the main effect of PRECISION BIAS was indeed significant ($\beta = -1.123$, $SE = 0.17$, $p < 0.001$, see Table 5.13). Results also showed that there was a PRECISION BIAS x NUMBER SIZE interaction such that the difference between the High and the Low Precision Bias conditions was significantly bigger for Small Numbers compared to Big Numbers ($F(2, 950.76) = 3.966$, $p < 0.05$, see Table 5.12). To further explore this interaction, I checked the effect of PRECISION BIAS on the three number sizes tested. The effect of PRECISION BIAS remained significant for Small Numbers ($\beta = -1.1836$, $SE = 0.2018$, $t = -5.865$, $p < 0.00001$) and Medium Numbers ($\beta = -0.7697$, $SE = 0.1793$, $t = -4.293$, $p < 0.0001$). For Big Numbers, the comparison did not reach significance but it did show a trend in the expected direction, with ratings being

higher in the High Precision Bias condition ($\beta = -0.3489$, $SE = 0.1909$, $t = -1.828$, $p > 0.7$). Finally, there also was a main effect of NUMBER SIZE ($F(2, 34.943) = 15.556$, $p < 0.001$), with all the comparisons among number sizes reaching significance (Small vs. Medium: $\beta = -0.7697$, $SE = 0.1793$, $t = -4.293$, $p < 0.0001$; Small vs. Big: $\beta = -1.3922$, $SE = 0.2257$, $t = -6.169$, $p < 0.000001$; Medium vs. Big: $\beta = -0.6974$, $SE = 0.1921$, $t = -3.631$, $p < 0.001$), such that the bigger the number, the lower the overall rating. Together, these results show that participants dispreferred to judge Small Numbers imprecisely and Big Numbers precisely.

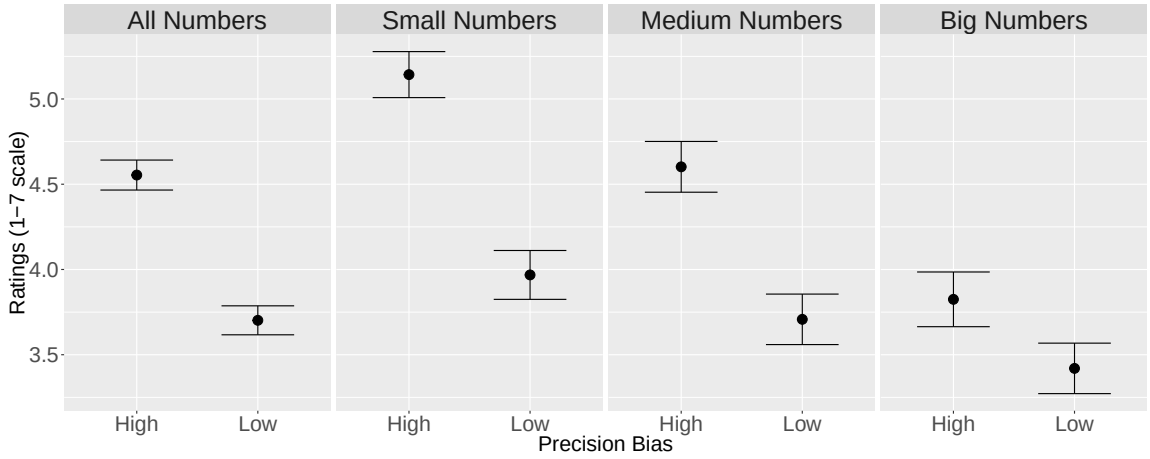


Figure 5.15: Precision Bias Ratings for the 24 experimental stimuli used in Experiment 6a. The graph plots ratings about the likelihood of the precise interpretation of the target numeral on a 1-7 scale, with 1 being *very unlikely* and 7 being *very likely*, for the two precision bias contexts and the three number sizes tested.

Predictor	Sum Sq	Mean Sq	NumDF	DenDF	F	p
NUMBER SIZE	69.887	34.943	2	22.22	15.556	< **0.001**
PLEVEL	83.328	83.328	1	60.10	37.095	< **0.001**
NUMBER SIZE : PBIAS	17.820	8.910	2	950.76	3.966	**0.01926**

Table 5.12: Analysis of Variance table for the precision ratings for the norming study of the experimental stimuli used in Experiments 6a, with F statistic, degrees of freedom, and p-values calculated using Satterthwaite's approximation.

The results of this norming study confirm that the experimental manipulations of Experiment 6a achieved the intended purpose: target numbers were judged to be more likely to be precise when preceded by a High Precision Bias context than when the same numeral was preceded by

Ratings ~ NUMBERSIZE * PLEVEL + (1 + NUMBERSIZE + PLEVEL | Subject) + (1 + NUMBERSIZE + PLEVEL | Item)				
Predictor	Coef β	SE(β)	t	p
Intercept	5.193	39.90	22.203	< 0.001
Medium Numbers	−0.709	0.19	−3.623	$<$ **0.001**
Big Numbers	−1.392	0.22	−6.169	$<$ **0.001**
PBias Low	−1.123	0.17	-6.288	$<$ **0.001**
Medium Numbers : PBias Low	0.402	0.22	1.763	> 0.08
Big Numbers : PBias Low	0.645	0.23	2.773	$<$ **0.01**

Table 5.13: Full mixed effects linear regression model predicting precision bias ratings for the 24 experimental stimuli used in Experiment 6a.

a Low Precision Bias context. Furthermore, the size of the numeral also played a role in guiding interpretations: the bigger the numeral, the less likely the precise interpretation was perceived to be. Taken as a whole, these results give us confidence about the interpretational preferences associated with the different conditions tested in Experiment 6a.

5.6.2 Predictions

Higher processing cost for the Low Precision Bias condition is expected if constructing an imprecise representation of the numeral as a result of integrating different contextual pragmatic cues follows the same mechanisms tracked by Experiments 3b and 4—where imprecise interpretations were triggered by the presence of a prenumeral approximator like *approximately* or *about*. Replicating the results of Experiments 3b and 4 would strengthen the conclusion that the processing of imprecision is more costly than the processing of precision, and would furthermore suggest that the effect of imprecision on processing is independent of the nature of the cues (i.e. semantic or pragmatic) that interpreters use to favor an imprecise interpretation over a precise one. If, on the other hand, the mechanisms underlying pragmatically driven imprecision calculation differ from the mechanisms underlying semantically driven imprecision calculation, the processing signature observed in Experiment 6a should differ from the patterns attested in Experiments 3b and 4. Such

findings would suggest that the greater processing cost observed for the imprecise conditions of Experiments 3b and 4 were the result of constructing an imprecise interpretation by semantically composing the slack regulator with the numeral.

5.6.3 Results

Accuracy Ratings

As in previous experiments, the proportion of correct responses was very high (the lowest mean was 83% in the Low Precision Bias condition for Big Numbers, see Figure 5.16). However, five participants were removed from data analysis, as they showed accuracy rates below 70%. A mixed effects model using PRECISION BIAS and NUMBER SIZE as fixed effects, and SUBJECT and ITEM as random effects was constructed to predict accuracy responses. Results show no significant difference across the condition means (all p's > 0.6), confirming that participants generally payed attention to the experimental task.

Reading Times

Next, I proceed to analyze the Reading Times (RTs). Visual inspection of the Raw Reading Times plotted in Figure 5.17 suggests that the experimental manipulations did not result in any relevant reading time differences in any of the two critical regions of interest, i.e. the Numeral and the Noun. This was confirmed by the results of two separate mixed effect models run on the data of the Numeral and the Noun regions respectively. Each of the models predicted Log residual reading times using PRECISION BIAS and NUMBER SIZE, and their interaction, as predictors and SUBJECT and ITEM as mixed effects. None of the predictors reached significance (all p's > 0.4, see Table 5.14 for the full model pertaining to the Numeral region, and Table 5.15 for the full list of model coefficients for the Noun region).

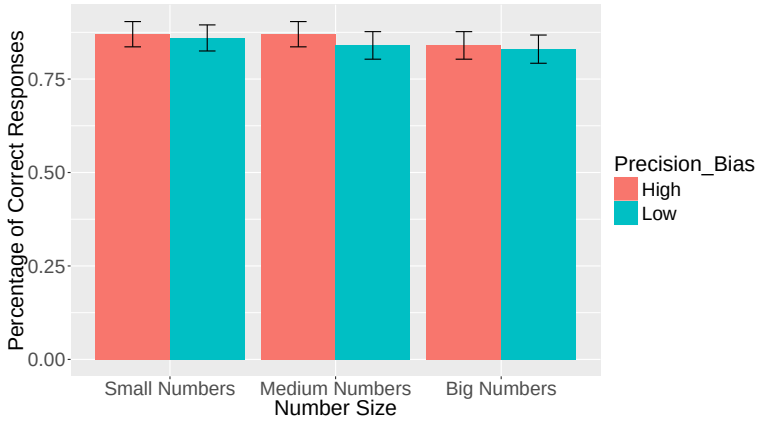


Figure 5.16: Accuracy Ratings for the six conditions tested in Experiment 6a.

Figure 5.17: Raw Reading Times for Experiment 6a broken down by Number Size. The horizontal axis plots the three regions of interest, i.e. the Numeral and the Noun regions, plus three words preceding and following the two regions of interest.

Figure 5.18: Log Residual Reading Times for Experiment 6a broken down by Number Size. The horizontal axis plots the two critical regions, i.e. the Numeral and the Noun regions, plus the three words preceding and following the critical regions of interest.

	Coef β	SE(β)	t	p
Intercept	−3.017	0.23	13.039	< 0.001
Medium Numbers	−0.025	0.037	−0.68	> 0.4
Big Numbers	−0.004	0.03	−0.134	> 0.8
Low Precision Bias	−0.039	0.035	−1.125	> 0.2
SPILLOVER$_{w_{-1}}$	0.378	0.046	8.061	< 0.001
SPILLOVER$_{w_{-2}}$	0.166	0.047	3.470	< 0.001
Medium Numbers : PBias Low	0.060	0.04	1.275	> 0.2
Big Numbers : PBias Low	0.021	0.04	0.455	> 0.6

Table 5.14: Full mixed effects linear regression model for the Number region of Experiment 6a.

	Coef β	SE(β)	t	p
Intercept	−2.673	0.23	−11.467	< 0.001
Medium Numbers	0.067	63.40	1.658	> 0.1
Big Numbers	0.074	0.03	1.856	> 0.06
PBias Low	0.025	0.03	0.665	> 0.5
SPILLOVER$_{w_{-1}}$	0.091	0.04	2.264	< 0.05
SPILLOVER$_{w_{-2}}$	0.392	0.04	8.018	< 0.001
Medium Numbers : Lower Precision Bias	−0.047	0.05	−0.903	> 0.3
Big Numbers : PBias Low	0.007	0.05	0.143	> 0.8

Table 5.15: Full mixed effects linear regression model for the Noun region of Experiment 6a.

5.6.4 Discussion

As was found in Experiment 5, Experiment 6a did not reveal any significant reading time differences between the High and the Low Precision Bias conditions, despite the fact that the offline judgement task used in the Norming Study (§5.6.1) revealed differences between these two conditions. More specifically, the Norming Study showed that the precision ratings belonging to the High Precision Bias condition were systematically higher than the ratings of the Low Precision Bias condition. The results also showed a significant interaction between PRECISION BIAS and NUMBER SIZE, such that the difference between the two Precision Bias conditions was attenuated for bigger numbers compared to the Small Numbers.

There exist at least four explanations for this mismatch between the offline and the online de-

pendent measures. First, it is possible that the perceived difference in the likelihood of the precise interpretation of the numeral detected in the Norming Study is unrelated to the online reading times as measured in Experiment 6b. If this was the case, the null effect found in Experiment 6a would indicate that there is no relation between the interpretation of the numeral and the local reading times. Second, it could be the case that such online processing correlate exists, but that self-paced reading is not a sensitive enough measure to detect such potential differences. Third, it is possible that the current results have been obscured by the fact that in the current version of the experiment the factor NUMBER SIZE is treated as a categorical variable (Small vs. Medium vs. Big) rather than as a continuous variable, which would provide us with finer grained distinctions. More importantly, the present results might be confounded by important individual differences among subjects. The current version of the experiment does not provide us with information about the actual interpretation of the numeral that the participant arrived at during reading. Being able to partition the results into different types of responders based on their individual precision bias ratings, rather than the two coarser-grained experimental conditions High vs. Low, might uncover more subtle differences in the patterns of online processing. Experiment 6b seeks to address these issues by treating the factor NUMBER SIZE as a continuous predictor of the reading times, and by collecting both precision bias ratings and reading times from all participants.

5.7 Experiment 6b

5.7.1 Methods

Participants

A total of 110 participants were recruited through Mechanical Turk and were payed $3 for their participation. Participants were all native speakers of English and were naïve to the purpose of the experiment. Data from 12 participants was discarded because they were older than 40 years old. (female = 47, mean = 29.13, range = 18-39).

Design

The experimental design closely follows that of Experiment 6a with some important modifications. Besides the 24 experimental items, I included a subexperiment with the purpose of detecting whether participants were paying attention to the task. The experiment consisted of the two conditions exemplified in (72). As in the experimental items, participants read a context paragraph followed by a target sentence.

(72) As the new student representative, Mary was receiving an increasing amount of emails from students concerned about their academic status. That morning, Mary received a message from an undergraduate who didn't know how many days there would be a hold placed on her account due to a late library book.

a. Mary replied that any student who **has** a late library book would have a hold placed on their account for 10 days.

b. Mary replied that any student who **have** a late library book would have a hold placed on their account for 10 days.

The target sentence always contained a numeral to maintain similarity with the experimental items. However, unlike the experimental items, the subexperiment target sentences always included a relative clause in subject position that either presented a number agreement error (72b), or was perfectly grammatical (72b). The critical region of interest, i.e. the verb inside the relative clause (bolded in example (72)), never coincided with the end of the sentence and was always followed by at least three more words.

Materials

The experimental items tested were the same as the ones used in Experiment 6a. Ten of the filler items used in Experiment 6a were substituted by the agreement subexperiment described in §5.7.1.

Procedure

The procedure parallels the one followed in Experiment 6a with the exception that participants received the instructions in written form and completed the experiment remotely, not in the laboratory. For each trial, participants first read the context paragraph. The target sentence was immediately presented in the following screen and was rendered in a moving window, word by word, self-paced reading fashion. The task used in this experiment was the same one used in the Norming Study reported in §5.6.1. Immediately after reading the target sentence, participants rated the likelihood of the precise interpretation of the target numeral by answering questions of the form in (73).

(73) How likely do you think it is that the nurse meant that exactly 10/100/1,000 patients were currently waitlisted?

Participants were provided a 1-7 scale with 1 being *very unlikely* and 7 being *very likely*. In order to answer the question, participants clicked on one of the seven numbers forming the scale. Neither the context nor the target sentence were available to the participants at the time of providing the precision rating. Besides collecting the precision ratings, reaction times, i.e. how long it took for participants to click on a number in the scale, were also recorded.

5.7.2 Predictions

Same as Experiment 6a.

5.7.3 Results

Accuracy Ratings

I first checked the accuracy ratings of the 10 filler trials for each of the participants. Five participants failed to reach the 70% threshold of correct responses and so their data was excluded from data analysis.

Agreement Subexperiment

Next, I proceed to report the results from the Agreement Subexperiment. The goal of the analysis of the subexperiment data is to establish a benchmark at which an effect that is independently known to exist, i.e. the subject-verb agreement violation effect (see Pearlmutter *et al.* 1999, and references therein), can be detected. The ultimate objective of this analysis is to more efficiently filter out participants who did not perform the experimental task appropriately.

To determine if readings times where higher for those trials containing a subject-verb agreement violation, I constructed a mixed effects linear regression model using two spillover regions and the factor AGREEMENT (Correct vs. Incorrect) as predictors. I also included SUBJECTS and ITEMS as random effects. The dependent variable was log-converted residual reading times. Results did not uncover any effects of Agreement type in neither the Agreement region ($\beta = 0.019$, $SE = 0.015$, $t = 1.252$, $p > 0.2$), the W1 region ($\beta = 0.009$, $SE = 0.017$, $t = 0.568$, $p > 0.5$) nor in the W2 region ($\beta = 0.035$, $SE = 0.022$, $t = 1.590$, $p > 0.1$). This null effect indicates that a significant proportion of the participants were not paying sufficient attention to the experimental task.

In order to detect those participants that mechanically pressed the space bar as fast as possible, without fully processing what they were reading, I identified those participants for whom more than 80% of the reading time data points was below 200 ms.[6] A total of 11 participants were discarded following this criterion. The same model described above was fit to the data of the remaining 82 participants. Again, both the Agreement region ($\beta = 0.004$, $SE = 0.0180$, $t = 0.226$, $p > 0.8$) and the W1 region ($\beta = 0.021$, $SE = 0.018$, $t = 1.185$, $p > 0.2$) failed to display any significant effects. However, the W2 region did show a trend in the relevant direction such that the condition with the incorrect agreement incurred higher reading times than the condition with the correct agreement ($\beta = 0.049$, $SE = 0.023$, $t = 2.077$, $p < 0.06$). To consolidate this trend, I removed those participants whose reading time means were two or more standard deviations away from the

6. This measure does not overlap with the accuracy ratings discussed above, since all the accuracy questions included in the filler trials targeted information that was presented in the context paragraph, not in the target sentence.

grand mean. A total of 3 participants were discarded following this second criterion, leaving a total of 79 participants (see Figure 5.19). As expected, the Agreement ($\beta = 0.002$, $SE = 0.018$, $t = 0.149$, $p > 0.8$) and the W1 regions ($\beta = 0.026$, $SE = 0.018$, $t = 1.422$, $p > 0.1$) were not significant, but the W2 region displayed a small but significant effect of Agreement ($F(1, 16.16) = 4.5491$, $p < 0.05$), with the incorrect agreement condition presenting significantly longer reading times than the correct agreement condition ($\beta = 0.053$, $SE = 0.025$, $t = 2.133$, $p < 0.05$, see Table 5.16).

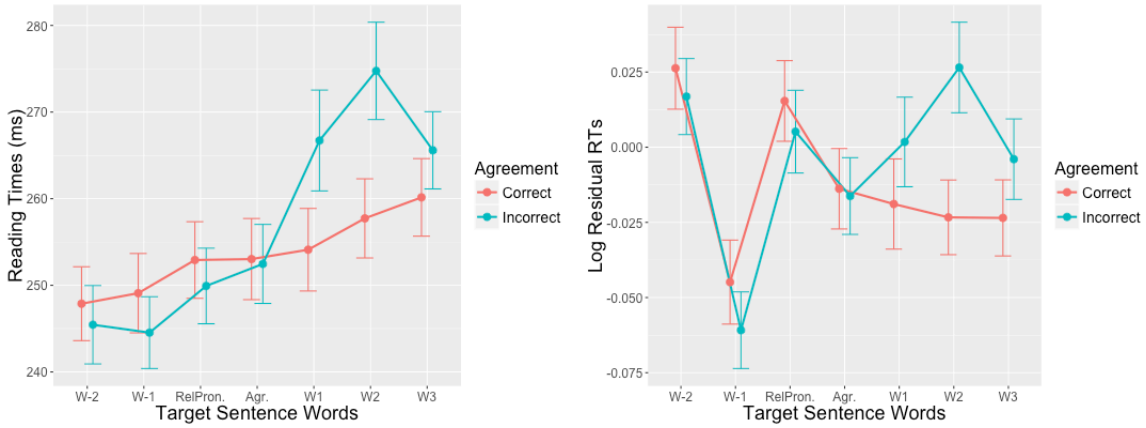


Figure 5.19: Left: Raw Reading Times for the subexperiment included in Experiment 6b from a total of 79 participants. The horizontal axis plots the relative pronoun and the verb of the relative clause, where agreement is resolved, plus two words preceding the relative pronoun and three words following the verb. Right: Log converted residual Reading Times for the target sentences of Experiment 6b after controlling for word length and word position. The horizontal axis contains the same regions plotted in the left panel.

Predictor	Coef β	SE(β)	t	p
Intercept	−1.317	0.171	−7.684	< 0.001
SPILLOVER$_{w_{-1}}$	0.066	0.026	2.511	< **0.0001**
SPILLOVER$_{w_{-2}}$	0.169	0.030	5.529	**0.0122**
Incorrect Agreement	0.053	0.025	2.133	< **0.0486**

Table 5.16: Full mixed effects linear regression model for the W2 region of the Agreement Subexperiment.

Despite the fact that the agreement violation effect is relatively small and delayed (as it does not reach significance until W2), we can now be more confident about the reliability of the pool of subjects that remained after the filtering procedures described in this section.

Ratings

The objective of this analysis is to determine whether the ratings obtained for the experimental trials of the current experiment match those of the Norming Study described in §5.6.1. Mean ratings for each of the 6 experimental conditions can be seen in Figure 5.20. Visual inspection of the plot suggests that the current ratings replicate those obtained in the Norming Study in the sense that the High Precision Bias condition was consistently rated higher than the Low Precision Bias condition in all number sizes. However, Figure 5.20 differs from the Norming Study's results in that there does not seem to be a PRECISION BIAS x NUMBER SIZE interaction.

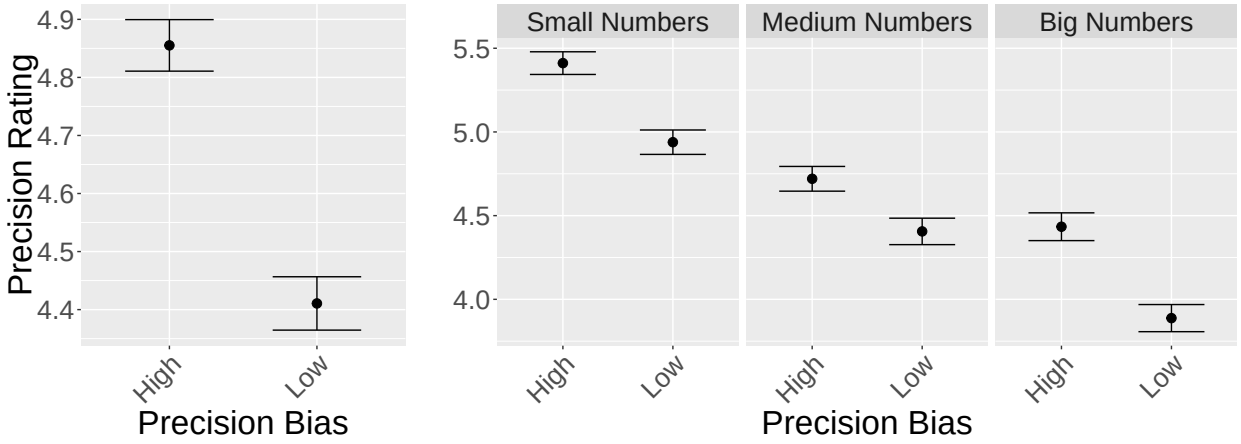


Figure 5.20: Precision Bias Ratings for the 24 experimental stimuli used in Experiment 6b. The graph plots ratings about the likelihood of the precise interpretation of the target numeral on a 1-7 scale, with 1 being *very unlikely* and 7 being *very likely*, for the two precision bias contexts and the three number sizes tested.

A mixed effects model using PRECISION BIAS, NUMBER SIZE and its interaction as predictors, and SUBJECTS and ITEMS as random effects was fit to the precision ratings data of Experiment 6b. Results confirmed the existence of a main effect of PRECISION BIAS ($F(1, 40.28) = 17.59$, $p < 0.001$, see Table 5.17), with the High Precision Bias condition receiving higher ratings than the Low Precision Bias condition ($\beta = -0.511$, $SE = 0.149$, $t = -3.409$, $p < 0.01$, see Table 5.18). The effect remained significant when the three number sizes were tested separately for both Small Numbers ($\beta = -0.5071$, $SE = 0.1816$, $t = -2.793$, $p < 0.01$) and Big Numbers ($\beta = -0.5745$, $SE =$

0.131, t = −4.37, $p < 0.0001$), whereas the comparison only showed a statistical trend for Medium Numbers ($\beta = -0.3378$, $SE = 0.1782$, t = −1.896, $p = 0.06$).

	Sum Sq	Mean Sq	NumDF	DenDF	*F*	*p*
NUMBER SIZE	42.164	21.0818	2	61.299	16.1578	< **0.0001**
PBIAS	22.942	22.9420	1	40.289	17.5836	**0.0001465**
NUMBER SIZE : PBIAS	3.879	1.9393	2	73.268	1.4863	0.2329438

Table 5.17: Analysis of Variance table for the precision ratings of Experiments 6b, with F statistic, degrees of freedom, and *p*-values calculated using Satterthwaite's approximation.

Ratings ~ PLEVEL * SCALE + (1 + PLEVEL * SCALE | Subject) + (1 + PLEVEL + SCALE | Item)				
Predictor	Coef β	SE(β)	t	*p*
Intercept	5.44909	0.20807	26.189	< 0.0001
Medium Numbers	−0.74144	0.15500	−4.784	< **0.0001**
Big Numbers	−1.02760	0.20288	−5.065	< **0.0001**
PBias Low	−0.51091	0.14986	−3.409	**0.00103**
PBias Low : Medium Numbers	0.19289	0.15857	1.216	0.22765
PBias Low : Big Numbers	−0.06044	0.15616	−0.387	0.69986

Table 5.18: Full mixed effects linear regression model predicting precision ratings of Experiment 6b, using the categorical predictor NUMBERSIZE.

Ratings ~ LOGNUMCENT * PLEVEL + (1 + LOGNUMCENT * PLEVEL | Subject) + (1 + LOGNUMCENT * PLEVEL | Item)				
Predictor	Coef β	SE(β)	t	*p*
Intercept	4.820473	0.226762	21.258	< 0.0001
LogNum	−0.229944	0.050807	−4.526	< **0.0001**
PBias Low	−0.427739	0.116105	−3.684	**0.000674**
LogNum : PBias Low	−0.007308	0.045828	−0.159	0.874057

Table 5.19: Full mixed effects linear regression model for the precision ratings of Experiment 6b using the continuous centered predictor LogNum.

Results also revealed a main effect of NUMBER SIZE ($F(2, 61.299) = 16.1578$, $p < 0.0001$, see Table 5.17), such that the ratings were increasingly lower for bigger numbers (Small vs. Medium: $\beta = -0.741$, $SE = 0.155$, $t = -4.784$, $p < 0.00001$; Medium vs. Big: $\beta = -0.2862$, $SE = 0.140$, $t = -2.043$, $p < 0.05$). The effect of NUMBER SIZE remained significant when the categorical predictor

NUMBER SIZE was substituted with a continuous centered predictor consisting of the log converted numerals used in the target sentence ($\beta = -0.229$, $SE = 0.050$, $p < 0.0001$, see Table 5.20). Finally, the interaction of the two factors tested was not significant in either of the two analyses (all p's > 0.2).

Reading Times

As in the previous self-paced reading studies reported in this chapter, observations above 1000 ms and below 100 ms were removed from data analysis. In Figures 5.21 and 5.22, I present the reading time data obtained for experiment 6b.

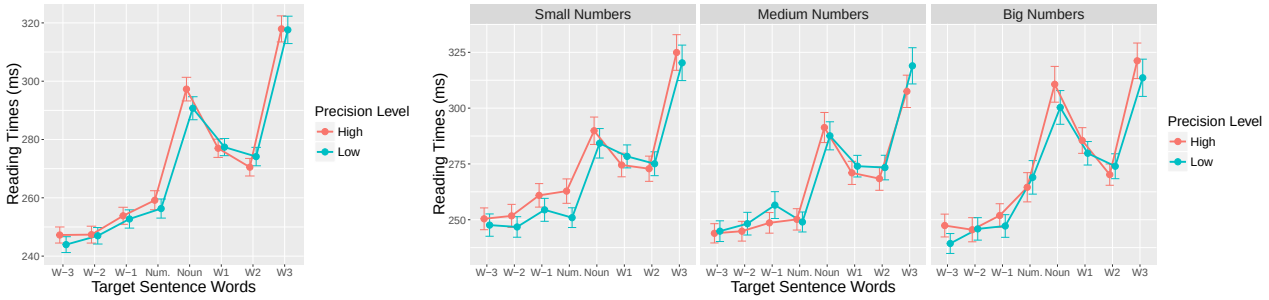


Figure 5.21: Raw Reading Times for Experiment 6b broken down by Number Size. The horizontal axis plots the two regions of interest, i.e. the Numeral and the Noun regions, plus three words preceding and following the two regions of interest.

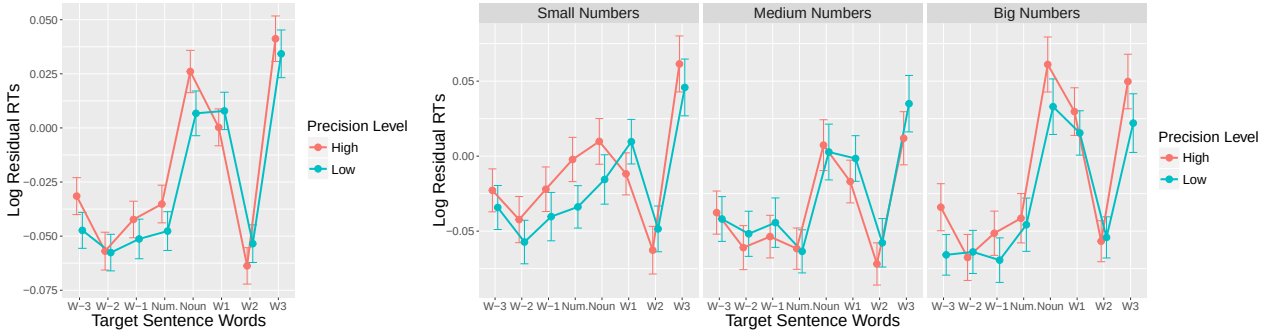


Figure 5.22: Log Residual Reading Times for Experiment 6b broken down by Number Size after controlling for word length and word position. The horizontal axis plots the two regions of interest, i.e. the Numeral and the Noun regions, plus three words preceding and following the two regions of interest.

Visual inspection of the plots, suggests that there were no relevant differences among the two precision bias conditions. In order to confirm that there were no significant effects, I fit the same model constructed for the analysis of the reading times of experiment 6a to the data pertaining to the Number and the Noun regions of Experiment 6b. No effects of PRECISION BIAS or interactions were detected at the Numeral region (all p's > 0.1) or at the Noun region (all p's > 0.2). I also substituted the categorical predictor NUMBER SIZE by the continuous predictor LOGNUM to predict the reading times. Again, results showed no significant differences in the Number region (all p's > 0.2) or the Noun region (all p's > 0.6).

Reaction Times

In this section, I analyze the second online measure collected in the current experiment, namely the response times, i.e. how long it took for participants to rate the likelihood of the precise interpretation. Figure 5.23 shows the reaction times for each of the seven points in the perceived precision level scale after trimming observations below 1000 ms and above 10000 ms.

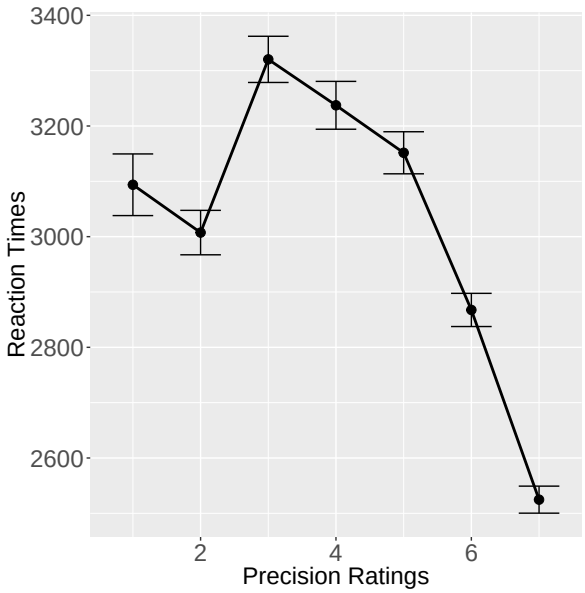


Figure 5.23: Reaction Times in ms (vertical axis) for the rating task included in Experiment 6b. The horizontal axis plots the 7 points forming the precision scale provided to the participants in order to judge the likelihood of the precise interpretation of the target numeral.

I constructed a model to predict the Reaction Times of the judgement responses from the RATINGS and LOGNUM, as well as their interaction. The model also included mixed effects for SUBJECT and ITEM. Results show that the ratings were a strong predictor of the Reaction Times ($\beta = -246.8$, $SE = 44.29$, $t = -5.572$, $p < 0.0001$, see Table 5.20), such that participants tended to be faster in judging the round numeral in those cases in which they assumed the speaker intended the number to be interpreted using a high precision threshold. No other significant effects were detected.

REACTIMES ~ LOGNUM * RATINGS + (1+ LOGNUM * REACTIMES | Subjects) + (1 + LOGNUM * REACTIMES | Items)				
Predictor	Coef β	SE(β)	t	p
Intercept	3182.94	129.45	24.588	< 0.0001
LogNum	38.69	33.32	1.161	0.249
Ratings	−246.80	44.29	−5.572	< **0.0001**
LogNum : Ratings	16.22	15.05	1.078	0.286

Table 5.20: Full mixed effects linear regression model for the Reaction Times of the judgement task included in Experiment 6b.

Post-hoc Analyses

A close inspection of Figure 5.23 reveals that the low scale points (1-2), corresponding to the imprecise interpretation of the numeral, display slower reaction times than the high scale points (6-7), corresponding to the precise interpretation. This Response Time asymmetry reveals that it took longer for participants to provide a judgement when they were confident that the speaker intended the numeral to be interpreted imprecisely as opposed to precisely. One possibility is that this asymmetry was not due to differences in the perceived precision threshold (i.e., high vs. low), but rather resulted from response bias effects. This is, it is possible that it was easier for participants to provide judgements for higher ratings in the scale than for lower ones. In order to rule out this possibility, I further explored the reading times of trials that participants judged to be the most extreme with respect to the perceived precision threshold. This corresponded to trials that were

judged with a 1 (i.e., trials that favored an imprecise interpretation) and with a 7 (i.e. trials that favored a precise interpretation, see Figure 5.24). This subset of the trials constituted 29.8% of the experimental data. As observed in the plot 5.24, the most extreme scale points do show a difference between the processing of (im)precision, with approximate interpretations incurring longer reading times than precise interpretations starting at the Number region. However, the difference between the reading times did not become significant until W2, the second postcritical region ($\beta = -0.07183$, $SE = 0.02961$, $t = -2.426$, $p < 0.03$).

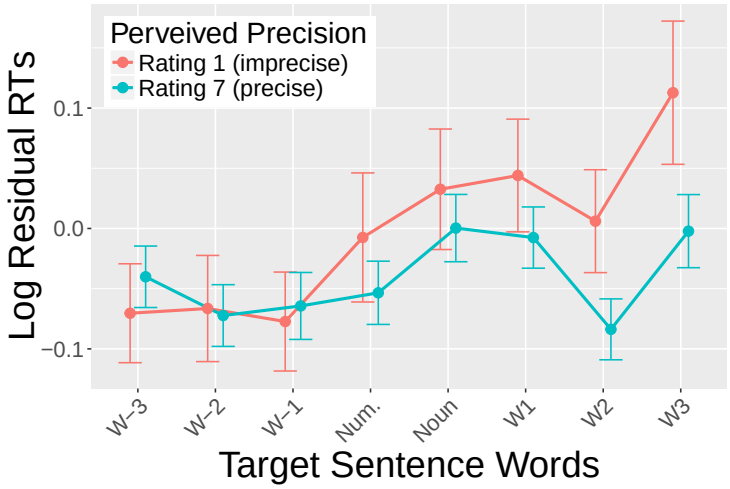


Figure 5.24: Reading Times as as function of Perceived Precision Level for the two most extreme ratings provided by participants. Rating 1 corresponds to trials for which participants were confident about the precise interpretation of the numeral. Rating 7 corresponds to trials for which participants were confident about the precise interpretation.

The fact that the effect in the reading times appears so late suggests that the task component, i.e. making judgements about precision, might play a role in the interpretation of the current results. It is possible that throughout the experiment, participants learned that after each target sentence they had to perform a precision judgement, and that this anticipation is what led to the reading time differences rather than the processing of the numeral itself.

To determine whether the reading time differences persisted beyond the trials rated more extremely, I also obtained mean precision ratings for each of the sentence/context combinations tested

(n=144). The first and fourth quintiles were selected (a total of 56 sentences and context pairs, 40% of the data). The sentences pertaining to the first and fourth quintiles were labeled based on their precision type, i.e. Low vs. High Precision respectively (see Figure 5.25). A Mixed Effects model using PRECISION LEVEL to predict the Log Residual Times was fit to the data pertaining to the Numeral and Noun regions. Results were not significant in the Numeral region ($p > 0.9$), but did show a significant processing penalty for Low Precision sentences in the Noun region, such that target sentences that were interpreted imprecisely incurred higher reading times in the noun region compared to target sentences containing numerals interpreted precisely ($\beta = 0.05978$, $SE = 0.0244$, $t = 2.45$, $p < 0.03$, see Table 5.21).

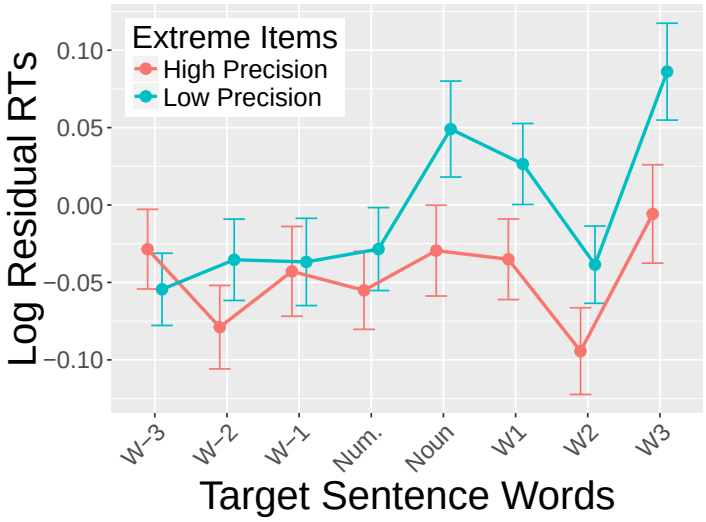


Figure 5.25: Reading times for the target sentences of Experiment 6b that received the 20% highest and lowest precision ratings on average (40% of the total amount of collected data). The red line represents reading times corresponding to target sentence/context pairs on the top quintile of the mean precision ratings (precise interpretation), whereas the blue line represents reading times for target sentence/context pairs on the bottom quintile of the mean precision ratings (imprecise interpretation).

The remaining 60% of the data, i.e. those target sentences that received intermediate precision ratings, was also analyzed following the same procedure (Figure 5.26). Analyses did not detect any significant effects of PRECISION LEVEL in either the Numeral or the Noun regions (all p's > 0.1).

In order to strengthen the reliability of this post-hoc analysis, and to effectively abstract away

Predictor	Sum Sq	Mean Sq	NumDF	DenDF	F	p
PLEVEL	0.36881	0.36881	1	21.72	6.002	**0.0228**
SPILLOVER$_{w_{-1}}$	2.07512	2.07512	1	699.99	33.769	< 0.00001
SPILLOVER$_{w_{-2}}$	0.9126	0.9126	1	697.88	14.851	< 0.00001

Table 5.21: Analysis of Variance table for the Log Residual Reading Times of the Noun Region of Experiment 6b. This analysis included only target sentences whose precision ratings were among the highest and lowest 20% respectively. The table shows the F-statistic, degrees of freedom, and p-values calculated using Satterthwaite's approximation.

LogRTRes ~ PLEVEL + (1 + PLEVEL | Subject) + (1 | Item)				
Predictor	Coef β	SE(β)	**t**	p
Intercept	-1.75367	0.1964	-8.929	< 0.00001
PLevel$_{Low}$	0.05978	0.0244	2.450	**0.022833**
SPILLOVER$_{w_{-1}}$	0.1853	0.03189	5.811	0.00001
SPILLOVER$_{w_{-2}}$	0.13028	0.0338	3.854	0.00001

Table 5.22: Full mixed effects linear regression model coefficients for the Noun Region of Experiment 6b. This analysis included only target sentences whose precision ratings were among the highest and lowest 20% respectively. The High Precision level was used as baseline for comparison

from potential task effect confounds, I applied the same analysis procedure to the data of Experiment 6a, which tested the same contexts and target sentences as Experiment 6b, but did not include a precision rating task. The goal of this last analysis is to determine whether the post-hoc analyses' results obtained for Experiment 6b can be replicated using the data of Experiment 6a. The reading times for the target sentences of Experiment 6a that received extreme and intermediate precision ratings in Experiment 6b can be found in Figures 5.27 and 5.28.

Visual observation of the plots reveals a qualitative difference between the High and the Low Precision Levels for both extreme and intermediate items, with the Low Precision Level trials incurring higher reading times in the Noun Region. This difference was however not significant for either of the two groups (all p's > 0.05). The lack of a significant effect raises the issue of the sparsity of the data collected for Experiment 6a. As seen in Table 5.23, which displays the amount of data submitted to statistical analysis in each of the post-hoc analyses discussed in this section, the

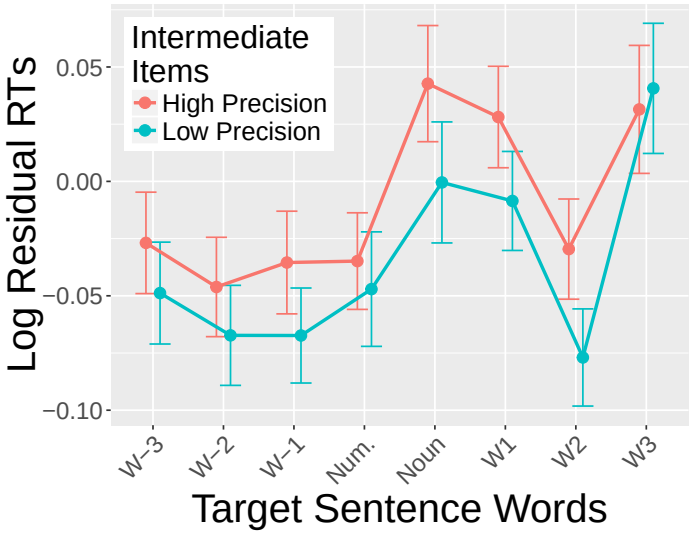


Figure 5.26: Reading times for the target sentences of Experiment 6b that received intermediate average precision ratings (60% of the total amount of collected data). Target sentences were further partitioned into target sentences/context pairs that received higher precision ratings within this group (red line, 50% of the data), and target sentences/context pairs that received the lower precision ratings within this group (blue line, 50% of the data).

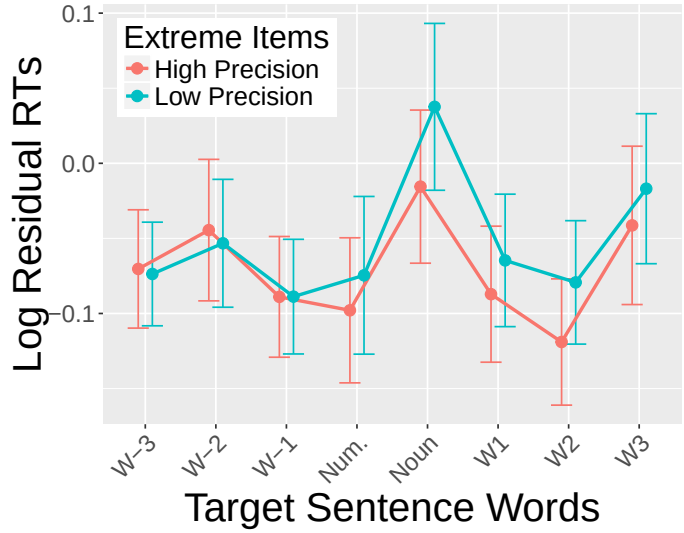


Figure 5.27: Reading times for the target sentences of Experiment 6a that received the 20% highest and lowest precision ratings on average in Experiment 6b (40% of the total amount of collected data). The red line represents reading times corresponding to target sentence/context pairs on the top quintile of the mean precision ratings (precise interpretation), whereas the blue line represents reading times for target sentence/context pairs on the bottom quintile of the mean precision ratings (imprecise interpretation)

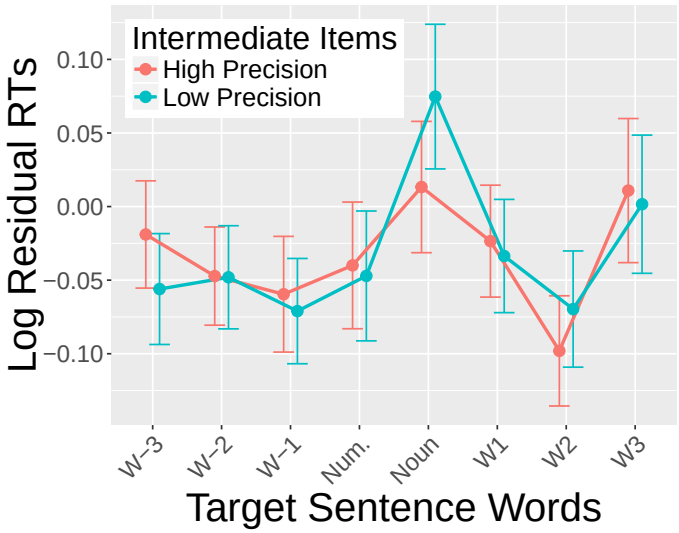


Figure 5.28: Reading times for the target sentences of Experiment 6a that received intermediate average precision ratings in Experiment 6b (60% of the total amount of collected data). Target sentences were further partitioned into target sentences/context pairs that received higher precision ratings within this group (red line, 50% of the data), and target sentences/context pairs that received the lower precision ratings within this group (blue line, 50% of the data).

data of Experiment 6b constitutes only around 32% of the amount of data collected for Experiment 6b. This suggests that the post-hoc analyses performed on the data of Experiment 6b were probably underpowered.

	Experiment 6b	**Experiment 6a**	Percentage
Extreme trials	5646	1877	40% of total
Intermediate trials	8494	2704	60% of total
Total trials	14140	4581	

Table 5.23: Amount of trials included in each of the post-hoc analyses described in this section for Experiments 6a and 6b.

To compensate for the sparsity of the data and provided that the two plots in Figures 5.27 and 5.28 are qualitatively very similar, I collapsed the Extreme and Intermediate items of Experiment 6a (see Figure 5.29) and repeated the same analysis procedure. Results showed a significant effect in the Noun Region (see Table 5.24) that paralleled the effect detected in the analysis of the extreme

items of Experiment 6b: a slow down in the reading times of the Low Precision sentences compared to the High Precision sentences ($\beta = 0.06919$, $SE = 0.0263$, t = 2.628, $p < 0.01$).

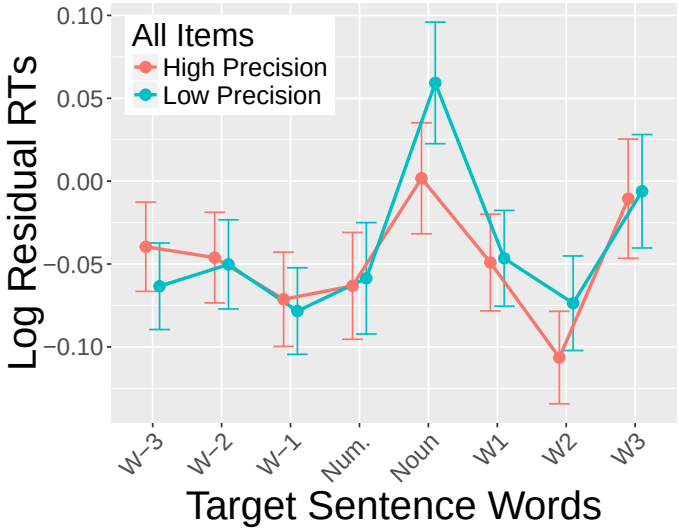


Figure 5.29: Reading times for the target sentences of Experiment 6a divided into High and Low Precision Level based on the precision ratings collected in Experiment 6b.

Predictor	Sum Sq	Mean Sq	NumDF	DenDF	F	p
PLEVEL	0.4593	0.4593	1	116.61	6.907	< **0.01**
SPILLOVER$_{w_{-1}}$	0.3564	0.3564	1	699.99	5.360	< 0.03
SPILLOVER$_{w_{-2}}$	4.0916	4.0916	1	550.78	61.540	< 0.00001

Table 5.24: Analysis of Variance table for Log Residual Reading Times pertaining to the Noun Region of Experiment 6a using the factor PLEVEL. The table shows the F-statistic, degrees of freedom, and p-values calculated using Satterthwaite's approximation.

LogRTRes ~ PLEVEL + (1 | Subject) + (1 | Item)				
Predictor	Coef β	SE(β)	**t**	p
Intercept	−2.62371	0.23426	−11.200	< 0.00001
PLevel$_{Low}$	0.06919	0.02633	2.628	< **0.01**
SPILLOVER$_{w_{-1}}$	0.09447	0.04081	2.315	0.02097
SPILLOVER$_{w_{-2}}$	0.38378	0.04892	7.845	0.00001

Table 5.25: Full mixed effects linear regression model coefficients for the Noun Region of Experiment 6a using PLEVEL as predictor and the High Precision level as baseline for comparison.

5.7.4 Discussion

Precision Threshold Ratings

The precision threshold ratings obtained in Experiment 6b show that participants were sensitive to the original manipulation (High vs. Low Precision Bias contexts), as they judged the High Precision Bias Contexts higher than the Low Precision Bias Context. However, the ratings obtained in Experiment 6b differ from the Norming Study's results (§5.6.1) in that no interaction between PRECISION BIAS and NUMBER SIZE was attested, although there was a main effect of NUMBER SIZE, with bigger numbers being judged more imprecisely than lower numbers. Taken together, these results suggest that the interaction detected in the Norming Study was not as stable as the main effect of PRECISION BIAS.

Reading and Reaction Times

Despite the existence of a main effect of PRECISION BIAS in the ratings, the Reading Rimes did not show any distinction between the High and the Low Precision contexts tested in Experiment 6b. However, when the two online measures collected in Experiment 6b were analyzed as a function of the precision ratings, similar patterns of results emerged for the two measures. The Reaction Times showed that participants were faster at judging trials in which they were certain that the speaker intended a *high* precision threshold (scale points 7), than trials in which participants were certain that the speaker intended a *low* precision threshold (scale points 1). However, the difference between the two groups only reached significance much later in the second post-critical region (W2). Together, these findings raise the issue of potential task-effect confounds. It is possible that participants were biased to select higher ratings in the precision rating scale, causing Reaction Times to be faster for responses corresponding to the higher end of the scale compared to responses corresponding to the lower end of the scale. Similarly, if participants learned throughout the experiment that after each trial they would be interrogated about the precision threshold of the target sentence,

the effect observed in the Reading Times of region W2 could be interpreted as resulting from the anticipation of the judgement task. This would explain why the effect was so delayed, and would suggest that the effect was not driven by the processing of the numeral, but was rather a by-product of the experimental task. Such result would be interesting in itself, but would not speak to the research question at hand.

When the reading time data of Experiment 6b was analyzed using the *mean precision ratings* of each item/condition pair, the penalty for imprecision in the Reading Times remained significant when the trials corresponding to item/condition pairs with the most extreme mean precision ratings were analyzed (40% of the data). The remaining data, corresponding to item/condition pairs that received intermediate precision ratings (60%), did not show any significant differences. Importantly, the effect detected for the extreme subset of the data appeared in the Noun Region, as would be expected if the slow down in the Reading Times was driven by the processing of the numeral. However, the question still remains of whether the judgement task could still be responsible for this pattern of results, since the effect was only present for the item/condition pairs that received the most extreme precision ratings on average. This issue can be settled by taking into consideration the results of the post-hoc analysis of Experiment 6a, which did not include a judgement task. As discussed in §5.7.3, Experiment 6a did not display any qualitative differences between the item/condition pairs that were judged to be more extreme vs. intermediate in Experiment 6b. However, when these two groups were collapsed, there was a clear penalty for the Low Precision group (lower mean ratings) compared to the High Precision group (higher mean ratings). Again, this penalty appeared in the Noun Region. The fact that we were still able to detect a processing penalty for imprecise interpretations in the Noun region in the absence of a judgement task confirms that this extra processing cost must have been driven by the interpretation of the imprecise numeral.

The current discussion was mostly based on post-hoc analyses and should therefore be taken with caution. However, given that all the post-hoc analyses discussed in §5.7.3 consistently showed

a processing penalty associated with imprecise interpretations, I would like to argue that the current findings provide further evidence for the view that the processing of imprecise interpretations of numerals is costlier than the processing of precise interpretations.

5.8 General Discussion

The sequence of experiments presented in this chapter had the goal of determining whether the processing of imprecise interpretations of round numbers is facilitated compared to the processing of their precise counterparts. The results described in this chapter have not provided any evidence supporting a processing advantage linked to imprecision calculation. On the contrary, all the significant differences detected in experiments 3 through 6 consistently showed a processing penalty for imprecise interpretations. I therefore conclude that imprecision calculation comes at a processing cost, not an advantage, at least for the comprehender.

The current results also bring up other important questions about the online processing of imprecise meanings. One of the objectives of this series of studies was to determine whether the online computation of imprecise meanings displayed the same processing signature when the comprehenders made use of cues of different nature—linguistics vs. pragmatic—, to construct and arrive at the imprecise representations. In experiments 3b and 4, participants received explicit linguistic cues, i.e. a Slack Regulator, that overtly signaled to them the precision threshold that was intended by the speaker. In this case, the construction of the imprecise representation of the round number interacted with the online semantic composition of the sentence, as the Slack Regulator directly modified the numeral. The second type of cue was tested in Experiment 6, and consisted of pragmatic manipulations about the context and goals of the conversation. Unlike the cues used in Experiments 3b and 4, the experimental manipulations used in Experiment 6 involved only pragmatic reasoning and did not interact with semantic composition. The results showed that both mechanism of signaling (im)precision slow down the processing of imprecise numerals, regardless of what type of knowledge, semantic or pragmatic is used by the comprehender to favor an impre-

cise interpretation. The fact that no processing facilitation was found for imprecise interpretations in the series of studies presented above does not rule out the possibility that imprecision might lead to some other sort of cognitive gain. However, until such empirical evidence is provided, the claim that imprecision is less costly than precision for the comprehender remains unsubstantiated. The current results suggest in fact otherwise: when cost is measured as a function of Reading Times, imprecision *is* costlier than precision.

5.9 Conclusion

In this chapter, I have presented a series of self-paced reading experiments that have investigated the processing of (im)precise interpretations of round numerals. Experiments 3b and 4, have investigated the processing of (im)precise interpretations induced by a slack regulator such us *about* or *exactly*. Experiments 6a and 6b, on the other hand, explored the processing of (im)precise interpretations that resulted from pragmatic reasoning about the goals of the conversation. To the extent that any significant differences were detected in these experiments, the direction of the effect was systematically the same: imprecise interpretations of round numbers incurred higher Reading Times than precise interpretations. These findings constitute the first empirical evidence that the processing of imprecision comes at a cost for the comprehener. Moreover, they suggest that it is the processing of imprecision itself what underlies the observed cost, as differences were detected when imprecision was explicitly signaled via an operator that interacts with semantic composition (i.e. a slack regulator) and when indirect pragmatic cues (i.e. manipulations about conversational goals) were used to bias comprehenders towards an (im)precise interpretation.

CHAPTER 6

CONCLUSION

6.1 Summary of Results and Implications

The work presented in this dissertation advances our understanding of the interactions between meaning and context in several respects. The results of the experiments presented in Chapter 3 provided further evidence that vagueness and imprecision should not be treated as a single type of context-sensitivity. This conclusion has implications for the theory of the relative vs. absolute distinction within the class of gradable adjectives. As was argued in Chapter 3, imprecision is a pragmatic phenomenon and should therefore not be encoded directly in the lexical representation of absolute adjectives.

The second half of this dissertation has focused on the processing of imprecision in the numeral domain. The results of the sequence of experiments presented in Chapter 5 show imprecise interpretations of round numerals incur higher processing cost than precise interpretations, regardless of the type of contextual cues, i.e. semantic or pragmatic, that comprehenders capitalize on to favor an imprecise interpretation over a precise one.

6.2 Future Directions

The next natural steps to continue the research started in this dissertation can be divided in two lines of work. First, in Chapter 3 it has been argued that within the class of AAs, only MaxAAs are subject to context-sensitive interpretations due to imprecision. I have argued that while there are context-sensitive *uses* of MinAAs, the truth-conditional meaning of MinAAs does not license context-sensitive interpretations of such adjectives. The lack of context-sensitivity found in all the experiments discussed in this dissertation for MinAAs supports this view. However, the strongest piece of evidence to validate this claim would come from data that shows that there exists an asymmetry between comprehension and production of context-sensitive utterances of MinAAs, such

that the production of MinAAs show context-sensitivity effects, while its interpretation does not. I hope to be able to provide such evidence in future research.

In the second part of this dissertation, I have provided the first empirical evidence that the processing of imprecise numerals incurs higher processing cost compared to the processing of precise interpretations. However, one aspect that has not been addressed in this work is whether there exist connections between the size of the pragmatic slack adopted for the interpretation of the numeral and the magnitude of the imprecision penalty observed in the Reading Tomes, or whether there exist interesting interactions between the size of the pragmatic slack tolerated in a context and the size of the numeral itself. These are all questions that should be addressed in further research.

A final important issue raised by the second part of this dissertation is determining what mechanisms are responsible for the extra processing cost identified for imprecise interpretations of numerals. I see at least two ways to approach this question. One possibility is that imprecise interpretations of numerals involve coercion of a default precise interpretations into a fuzzy range construal. There exists evidence that coercion involves cost in the verbal domain (Pickering *et al.* 2005). Therefore, to the extent that coercion is required in order to arrive at an imprecise interpretation of a round number, the additional computations involved in coercing a precise representation into an imprecise one could be driving the effects detected in Experiments 3-6. A second option is that the lexical representation of numerals is underspecified, as has been claimed to be the case for polysemous words (see Eddington & Tokowicz (2015) for a recent literature review on the topic), and that resolving imprecise meanings is harder than resolving precise interpretations, thus leading to more cost in the former case. Although these questions were beyond the scope of this dissertation, their answer will greatly contribute to the interpretation of the results reported in Chapter 5.

Finally, imprecision is known to be a pervasive cross-categorical phenomenon that affects not only the categories discussed in this dissertation, MaxAAs and numerals, but also prepositions, nouns or quantifiers, to mention a few. Little is known about whether certain grammatical cat-

egories present restrictions with respect to imprecision calculation or not. Gaining a better understanding of whether imprecision involves additional processing cost in other grammatical categories beyond the numeral domain is necessary in order to understand whether the mechanisms underlying imprecision calculation are also cross-categorical, or whether imprecision arises via different mechanisms depending on the grammatical constraints imposed by different categories.

REFERENCES

Allopenna, P. D., James S. Magnuson, & Michael K. Tanenhaus. 1998. Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language* 419–439.

Altmann, Gerry, & Mark Steedman. 1988. Interaction with context during human sentence processing. *Cognition* 30.191–238.

Alxatib, Sam, & Jeff Pelletier. 2009. On the psychology of truth-gaps. In *International Workshop on Vagueness in Communication*, 13–36. Springer.

Anderson, Curt. 2014. Approximation of complex cardinals using *some*. In *Proceedings of the Western Conference on Linguistics (WECOL) 2013*, ed. by Claire Renaud, Carla Ghanem, Verónica González López, & Kat Pruitt, 131–143.

——. 2015. Numerical approximation using *some*. In *Proceedings of Sinn und Bedeutung (SuB) 19*, ed. by Eva Csipak & Hedde Zeijlstra, 54–70.

Aparicio, Helena, Ming Xiang, & Christopher Kennedy. 2015. Processing gradable adjectives in context: a Visual World study. In *Proceedings of SALT 25*.

Arts, Anja, Alfons Maes, Keo Noorman, & Carel Jansen. 2011. Overspecification facilitates object identification. *Journal of Pragmatics* 43.361–374.

Baayen, R. H., D. J. Davidson, & D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59.390–412.

Barker, Chrisr. 2002. The dynamics of vaguness. *Linguistics and Philosophy* 25.1–36.

Barr, Dale J., Roger Levy, Christoph Scheepers, & Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing. Keep it maximal. *Journal of Memory and Language* 68.

Bartsch, Renate, & Theo Vennemann. 1973. *Semantic structures: A study in the relation between syntax and semantics*. Bartsch, Renate and Vennemann, Theo.

Bastiaanse, Harald. 2011. The rationality of round interpretation. In *Vagueness in communication. International workshop, ViC 2009 held as part of ESSLLI 2009, Bordeaux, France, July 20-24, 2009. Revised Selected Papers*, ed. by Rick Nouwen, Robert van Rooij, Uli Sauerland, & Hans-Christian Schmitz, volume 6517 of *Lecture Notes in Computer Science*, 37–50. Berlin-Heilderberg: Springer.

Belke, Eva, & Antje S. Meyer. 2002. Tracking the time course of multidimensional stimulus discrimination: Analyses of viewing patterns and processing times during "same"–"different" decisions. *European Journal of Cognitive Psychology* 14.237–266.

Bergen, Leon, Noah Goodman, & Roger Levy. 2012. That's what she (could have) said: How alternative utterances affect language use. In *Proceedings of the Cognitive Science Society*, volume 34.

Bierwisch, M. 1989. The semantics of gradation. In *Grammatical Structure and Conceptual Interpretation*, ed. by M. Bierwisch & E. Langs, 71–261. Springer-Verlag.

Bolinger, Dwight. 1972. *Degree words*, volume 53. Walter de Gruyter.

Bonini, Nicolao, Daniel Osherson, Riccardo Viale, & Timothy Williamson. 1999. On the psychology of vague predicates. *Mind & Language* 14.377–393.

Burnett, Heather. 2014. A delineation solution to the puzzles of absolute adjectives. *Linguistics and Philosophy* 37.1–39.

Burns, Lynda. 2012. *Vagueness: An investigation into natural languages and the sorites paradox*, volume 4. Springer Science & Business Media.

Chambers, Craig G, Michael K Tanenhaus, Kathleen M Eberhard, Hana Filip, & Greg N Carlson. 2002. Circumscribing referential domains during real-time language comprehension. *Journal of Memory and Language* 47.30–49.

Cooper, R. M. 1974. The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology* 6.84–107.

Crain, S., & Mark Steedman. 1985. On not being led up the garden path: the use of context by the psychological parser. In *Natural Language Parsing*, ed. by D. Dowty, L. Kartunnen, & A. Zwicky. Cambridge, MA: Cambridge University Press.

Creswell, M. J. 1976. The semantics of degree. In *Montague Grammar*, ed. by B. Partee, 261–292. New York: Academic Press.

Cruse, D Alan. 1980. Antonyms and gradable complementaries. *Perspektiven der lexikalischen semantik: Beitraäge zum wuppertaler semantikkolloquium vom* 2–3.

Cummins, Chris, Uli Sauerland, & Stephanie Solt. 2012. Granularity and scalar implicature in numerical expressions. *Linguistics and Philosophy* 35.35–135.

Cummins, Christopher. 2018. modified fractions, granularity and scale structure. In *The semantics of gradability, vagueness, and scale structure: experimental perspectives*, ed. by Galit W. Sassoon, Louise McNally, & Elena Castroviejo-Miro. Springer.

Dahan, Delphine, James S. Magnuson, & Michael K. Tanenhaus. 2001. Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology* 42.317–367.

De Jaegher, Kris, & Robert van Rooij. 2009. Strategic vagueness, and appropriate contexts. In *Discussion Paper Series/Tjalling C. Koopmans Research Institute*, volume 9, 9–24. UU USE Tjalling C. Koopmans Research Institute.

Dehaene, Stanislas. 2003. The neural basis of the weber–fechner law: a logarithmic mental number line. *Trends in cognitive sciences* 7.2003.

——, & Jacques Mehler. 1992. Cross-linguistic regularities in the frequency of number words. *Cognition* 43.1–29.

Doetjes, J., C. Constantinescu, & Součková. 2009. A neo-kleinian approach to comparatives. In *Proceedings of Semantics and Linguistic Theory (SALT) 19*, ed. by D. Lutz E. Cormany, S. Ito, 124–141.

Eberhard, Kathleen M., Michael J. Spivey-Knowlton, Julie C. Sedivy, & Michael K. Tanenhaus. 1995. Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research* 24.409–436.

Eddington, Chelsea M, & Natasha Tokowicz. 2015. How meaning similarity influences ambiguous word processing: The current state of the literature. *Psychonomic bulletin & review* 22.13–37.

Ferson, Scott, Jason O'Rawe, Andrei Antonenko, Jack Siegrist, James Mickley, Christian C Luhmann, Kari Sentz, & Adam M Finkel. 2015. Natural language of uncertainty: numeric hedge words. *International Journal of Approximate Reasoning* 57.19–39.

Frank, Michael C., & Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336.998–998.

Frank, Michael C, & Noah D Goodman. 2014. Inferring word meanings by assuming that speakers are informative. *Cognitive psychology* 75.

Frazee, Joey, & David Beaver. 2010. Vagueness is rational under uncertainty. In *Logic, language and meaning*, 153–162. Springer.

Gibbs, Raymond W., & Gregory A. Bryant. 2008. Striving for optimal relevance when answering questions. *Cognition* 106.345–369.

Gibson, Edward. 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition* 68.1–76.

——. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain* 95–126.

Goodman, Noah D., & Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences* 20.818 – 829.

——, & Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science* 5.173–184.

Graff, Delia. 2000. Shifting sands: An interest-relative theory of vagueness. *Philosophical topics* 28.45–81.

Grano, Thomas. 2012. Mandarin hen and universal markedness in gradable adjectives. *Natural Language and Linguistic Theory* 30.513–565.

Grice, Herbert Paul. 1975. Logic and conversation. In *Syntax and Semantics Volume 3: Speech Acts*, ed. by P. Cole & J. Morgan. New York: Academic Press.

Griffin, Zenzi M, & Kathryn Bock. 2000. What the eyes say about speaking. *Psychological science* 11.274–279.

Grodner, D., & Julie C. Sedivy. 2011. The effect of speaker-specific information on pragmatic inferences. In *The processing and acquisition of reference*, ed. by N. Pearlmutter & E. Gibson, 239–272. Cambridge, MA: MIT Press.

Hale, John. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research* 32.101–123.

Hanna, Joy E., & Michael K. Tanenhaus. 2004. Pragmatic effects on reference resolution in a collaborative task: Evidence from eye movements. *Cognitive Science* 28.105–115.

——, ——, & John C. Trueswell. 2003. The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language* 49.43–61.

Heim, Irene. 2000. Degree operators and scope. In *Proceedings of Semantics and Linguistic Theory (SALT) 10*, ed. by B. Jackson & T. Matthews, 40–64.

Hobbs, Jerry R, & Vladik Kreinovich. 2006. Optimal choice of granularity in commonsense estimation: Why half-orders of magnitude? *International Journal of Intelligent Systems* 21.843–855.

Horn, Laurence R. 1984. Towards a new taxonomy of pragmatic inference: Q-based and r-based implicature. In *Meaning, form, and use in context: Linguistic applications*, ed. by D. Schiffring, 11–89. Washington D.C.: Georgetown University Press.

Huettig, Falk, Joost Rommers, & Antje Meyer. 2011. Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica* 137.151–171.

Izard, Véronique, & Stanislas Dehaene. 2008. Calibrating the mental number line. *Cognition* 106.1221–1247.

——, Pierre Pica, Elizabeth S Spelke, & Stanislas Dehaene. 2008. Exact equality and successor function: Two key concepts on the path towards understanding exact numbers. *Philosophical Psychology* 21.491–505.

Jäger, Gerhard. 2012. Game theory in semantics and pragmatics. In *Semantics: An international handbook of natural language meaning*, volume 3, 2487–2425. Mouton de Gruyter Berlin.

Jansen, C. J. M., & M. M. W. Pollmann. 2001. On round numbers: Pragmatic aspects of numerical expressions. *Journal of Quantitative Linguistics* 8.187–201.

Kamide, Yuki, Gerry TM Altmann, & Sarah L Haywood. 2003. The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language* 49.133–156.

Kamp, H. 1985. Two theories of adjectives. In *Formal Semantics of Natural Language*, ed. by E. Keenan, 123–155. Cambridge: Cambridge University Press.

Kao, Justine T., Jean Y. Wu, Leon Bergen, & Noah D. Goodman. 2014. Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences* 111.12002–12007.

Kennedy, Christopher. 1999. *Projecting the adjective: The syntax and semantics of gradability and comparison*. New York: Garland Press.

——. 2007. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy* 30.1–45.

——, & Louise McNally. 2005. Scale structure and the semantic typology of gradable predicates. *Language* 81.345–381.

——, & Louise McNally. 2010. Color, context, and compositionality. *Synthese* 174.79–98.

Kim, Christina S., Christine Gunlogson, Michael K. Tanenhaus, & Jeffrey T. Runner. 2015. Context-driven expectations about focus alternatives. *Cognition* 139.28–49.

Kingston, John, Joshua Levy, Amanda Rysling, & Adrian Staub. 2016. Eye movement evidence for an immediate Ganong Effect. *Journal of Experimental Psychology: Human Perception and Performance* 1–22.

Klecha, Peter. *Bridging the divide: Scalarity and Modality*. University of Chicago dissertation.

——. 2017. On unidirectionality in precisification. *Linguistics and Philosophy* .

Klein, Ewan. 1980. A semantics for positive and comparative adjectives. *Linguistics and Philosophy* 4.1–45.

Koolen, Ruud, A. Gatt, M. Goudbeek, & Krahmer E. 2011. Factors causing overspecification in definite descriptions. *Journal of Pragmatics* 43.3231–3250.

——, M. Goudbeek, & Krahmer E. 2013. The effect of scene variatio on the redundant use of color in definite reference. *Cognitive Science* 37.395–411.

Krifka, Manfred. 2002. Be brief and vague! and how bidirectional optimality theory allows for verbosity and precision. In *Sound and Systems. Studies in Structure and Change. A Festschrift for Theo Vennemann*, ed. by D. Restle & D. Zaefferer, 439–358. Berlin: Mouton de Gruyter.

——. 2007. Approximate interpretations of number words: A case of strategic communication. In *Cognitive Foundations of Interpretation*, ed. by Gerlof Bouma, Irene Krämer, & Joost Zwarts, 111–126. Amsterdam: Humboldt-Universität zu Berlin, Philosophische Fakultät II.

——. 2009. Approximate interpretations of number words: a case of strategic communication. In *Theory and evidence in semantics*, ed. by Erhard Hinrichs & John Nerbonne, 109–132. Stanford: CSLI Publications.

Lakoff, George. 1973. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic* 2.458–508.

Lasersohn, Peter. 1999. Pragmatic halos. *Language* 75.522–551.

Lassiter, Daniel, & Noah D. Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and Linguistic Theory (SALT) 23*, ed. by Todd Snider, 587–610, Ithaca, NY. CLC.

——, & ——. 2015. Adjectival vagueness in a Bayesian model of interpretation. *Synthese* 1–36.

Lauer, Sven. 2012. On the pragmatics of pragmatic slack. In *Proceedings of Sinn und Bedeutung 16*, ed. by Ana Aguilar Guevara, Anna Chernilovskaya, & Rick Nouwen, 389–401, MA. MIT Working Papers.

Leffel, Timothy, Ming Xiang, & Christopher Kennedy. 2016. Imprecision is Pragmatic: Evidence from referential processing. In *Proceedings of SALT 26*, 836–854.

Lefort, Sébastien, Marie-Jeanne Lesot, Elisabetta Zibetti, Charles Tijus, & Marcin Detyniecki. 2017. Interpretation of approximate numerical expressions: Computational model and empirical study. *International Journal of Approximate Reasoning* 193–209.

Levinson, Stephen. 2000. *Presumptive Meanings*. MA: MIT Press.

Levy, Roger. 2008. Expectation-based syntactic comprehension. *Cognition* 106.1126–1177.

Lewis, D. 1979. Scorekeeping in a language game. *Journal of Philosophical Logic* 8.339–359.

Lewis, Richard L, & Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science* 29.375–419.

Luce, Paul A., & David B. Pisoni. 1998. Recognizing spoken words: the neighborhood activation model. *Ear and Hearing* 19.1–36.

Ludlow, Peter. 1989. Implicit comparison classes. *Linguistics and Philosophy* 12.519–533.

Maes, Alfons, Anja Arts, & L. Noordman. 2004. Reference management in instructive discourse. *Discourse Process* 37.117–144.

Marslen-Wilson, William D., & Alan Welsh. 1978. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychologysychology* 10.29–63.

Mason, J. David, Alice F. Healy, & William R. Marmie. 1996. The effects of rounding on memory for numbers in addition problems. *Canadian Journal of Experimental Psychology* 50.320–323.

McNally, Louise. 2011. The relative role of property type and scale structure in explaining the bahavior of gradable adjectives. In *Vagueness in Communication*, ed. by Rick Nouwen, Robert van Rooij, Uli Sauerland, & Hans-Christian Schmitz, volume 6517 of *Lecture Notes in Computer Science*, 151–168. Berlin Heidelberg: Springer.

Meyer, Antje S, Astrid M Sleiderink, & Willem JM Levelt. 1998. Viewing and naming objects: Eye movements during noun phrase production. *Cognition* 66.B25–B33.

Morzycki, Marcin. 2012. Adjectival extremeness: Degree modification and contextually restricted scales. *Natural Language and Linguistic Theory* 30.567–609.

Nadig, Aparna S., & Julie C. Sedivy. 2002. Evidence of perspective-taking contraints in children's on-line reference resolution. *Psychological Science* 13.329–336.

Paraboni, Ivandré, Kees van Deemter, & Judith Masthoff. 2007. Generating referring expressions: Making referents easy to identify. *Computational Linguistics* 33.229–254.

Pearlmutter, Neal J., Susan M. Garnsey, & Kathryn Bock. 1999. Agreement processes in sentence comprehension. *Jouran of Memory and Language* 41.

Pechmann, Thomas. 1989. Incremental speech production and referential overspecification. *Linguistics* 27.89–110.

Peloquin, Benjamin, & Michael C Frank. 2016. Determining the alternatives for scalar implicatures. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, ed. by A. Papafragou, D. Grodner, D. Mirman, & J.C. Trueswell, 319–324.

Piantadosi, Steven T, Harry Tily, & Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition* 122.280–291.

Pickering, Martin J, Brian McElree, & Matthew J Traxler. 2005. The difficulty of coercion: A response to de almeida. *Brain and Language* 93.1–9.

Pinkal, Manfred. 1995. *Logic and lexicon: the semantics of the indefinite*. Dordrecht: Kluwer.

Pogue, Amanda, C. Kurumada, & Michael K. Tanenhaus. 2016. Talker-specific generalization of pragmatic inferences based on under- and over-informative prenominal adjective use. *Frontiers in Psychology* 6.

Potts, Christopher. 2008. Interpretive economy, schelling points, and evolutionary stability. *Manuscript, UMass Amherst* .

——, Daniel Lassiter, Roger Levy, & Michael C. Frank. 2016. Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of Semantics* 33.755–802.

Pyykkönen-Klauck, Pirita, & Matthew W. Crocker. 2016. Attention and eye movement metrics in visual world eye tracking. In *Visually situated language comprehension*, ed. by Pia Knoeferle, Pirita Pyykkönen-Klauck, & Matthew W. Crocker, 67–82. John Benjamins.

Qing, Ciyang, & Michael Franke. 2014. Gradable adjectives, vagueness, and optimal language use: a speaker-oriented model. In *Semantics and Linguistic Theory (SALT) 24*, ed. by Todd Snider, Sarah D'Antonio, & Mia Weigand, 23–41, Ithaca, NY. CLC.

Rayner, Keith. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin* 124.372–422.

Rotstein, Carmen, & Yoad Winter. 2004. Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers. *Natural Language Semantics* 12.259–288.

Rubio-Fernández, Paula. 2015. Redundancy is efficient –and effective, too. In *Paper presented at the XI Conference on Architectures and Mechanisms for Language Processing (AMLaP)*.

——. 2016. How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in Psychology* 7.

Sadock, Jerrold M. 1977. Truth and approximations. In *Annual Meeting of the Berkeley Linguistics Society*, volume 3, 430–439.

Sassoon, Galit, & Assaf Toledo, 2011. Absolute and relative adjectives and their comparison classes. Unpublished manuscript.

Sauerland, Uli, & Penka Stateva. 2007. Scalar vs. epistemic vagueness: Evidence from approximators. In *Proceedings of SALT 17*, ed. by T. Friedman & M Gibson, 228–245, Ithava, NY. Cornell University.

Scontras, G., & Michael Henry Tessler. 2017. Probabilistic language understanding: an introduction to the Rational Speech Act framework.

Sedivy, Julie C. 2003. Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research* 32.3–23.

—— 2004. Evaluating explanations for referential context effects: Evidence for Gricean meachanisms in online language interpretation. In *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions*, ed. by John C. Trueswell & Michael K. Tanenhaus, 345–364. Cambridge, MA: MIT Press.

——, Michael K. Tanenhaus, Craig G. Chambers, & Gregory N. Carlson. 1999. Achieving incremental semantic interpretation through contextual representation. *Cognition* 71.109–147.

Shannon, Claude E. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27.379–423.

Sigurd, Bengt. 1988. Round numbers. *Language in Society* 17.243–252.

Solt, Stephanie. 2015. Vagueness and imprecision: Empirical foundations. *Annu. Rev. Linguist.* 1.107–127.

——, Chris Cummins, & Marijan Palmović. 2017. The preference for approximation. *International Review of Pragmatics* 9.248–268.

——, & Nicole Gotzner. 2012. Experimenting with degree. In *Proceedings of SALT*, volume 22.

Sperber, Dan, & D. Wilson. 1995. *Relevance: Communicationa and cognition*. Oxford: Blackwell.

——, & Deirdre Wilson. 1985. Loose talk. In *Proceedings of the Aristotelian Society*, volume 86, 153–171.

Syrett, Kristen, Christopher Kennedy, & Jeffrey Lidz. 2009. Meaning and context in children's understanding of gradable adjectives. *Journal of Semantics* 27.1–35.

Tanenhaus, Michael K., Michael J. Spivey-Knowlton, Kathleen M. Eberhard, & Julie C. Sedivy. 1995. Integration of visual and linguistic information during spoken language comprehension. *Science* 268.1632–1634.

Tessler, Michael Henry, Michael Lopez-Brau, & Noah D Goodman. 2017. Warm (for winter): Comparison class understanding in vague language. *Proceedings of the Annual Meeting of the Cognitive Science Society* .

Toledo, Assaf, & Galit W. Sassoon. 2011. Absolute vs. relative adjectives – variance within vs. between individuals. In *Proceedings of SALT 21*, 135–154.

Trueswell, John C, Michael K Tanenhaus, & Susan M Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of memory and language* 33.285–318.

Unger, Peter. 1975. *Ignorance: a case for scepticism*. Oxford: Clarendon Press.

van der Henst, Jean-Baptiste, Laure Carles, & Dan Sperber. 2002. Truthfulness and relevance in telling the time. *Mind & Language* 17.457–466.

van Rooij, Robert. 2011. Vagueness and linguistics. In *Vagueness: A Guide*, ed. by G. Ronzitti, chapter 6, 123–179. Dordrecht: Springer.

von Setechow, A. 1984. Comparing semantic theories of comparison. *Journal of Semantics* 3.1–77.

Weber, Andrea, Bettina Braun, & Matthew W. Crocker. 2006. Finding referencts in timr: Eye-tracking evidence for the role of contrastive accents. *Language and Speech* 49.367–392.

Westerbeek, Hans, Ruud Koolen, & Alfons Maes. 2015. Stored object knowledge and the production of referring expressions: the case of color typicality. *Frontiers in Psychology* 6.

Wilson, D., & Dan Sperber. 2002. Truthfulness and relevance. *Mind* 111.583–632.

Wolter, Lynsey, Kristen Skovbroten Gorman, & Michael K. Tanenhaus. 2011. Scalar reference, contrast and discourse: Separating effects of linguistic discourse from availability of the referent. *Journal of Memory and Language* 65.299–317.

Yoon, Youngeun. 1996. Total and partial predicates and the weak and strong interpretations. *Natural Language Semantics* 4.217–236.

Zeevat, Henk. 2014. *Language production and interpretation: Linguistics meets cognition*. Brill.

Zipf, G. K. 1949. *Human behavior and the principle of least effocrt*. Cambridge: Addison-Wesley.

APPENDIX A
FULL MODELS FOR EXPERIMENT 1

	Coef β	SE(β)	z Value	p
Intercept	−2.570	0.326	−7.863	0.001
Condition	−0.1104	0.3603	−0.306	0.8

Table A.1: 0-100 ms window (ColAs).

	Coef β	SE(β)	z Value	p
Intercept	−2.736	0.307	−8.889	0.001
Condition	0.176	0.319	0.553	0.6

Table A.2: 100-200 ms window (ColAs).

	Coef β	SE(β)	z Value	p
Intercept	−2.569	0.279	−9.179	0.001
Condition	0.050	0.342	0.147	0.9

Table A.3: 200-300 ms window (ColAs).

	Coef β	SE(β)	z Value	p
Intercept	−1.843	0.215	−8.561	0.001
Condition	−0.285	0.288	−0.992	0.4

Table A.4: 300-400 ms window (ColAs).

	Coef β	SE(β)	z Value	p
Intercept	−1.304	0.235	−5.537	0.001
Condition	−0.301	0.276	−1.089	0.3

Table A.5: 400-500 ms window (ColAs).

	Coef β	SE(β)	z Value	p
Intercept	−0.7506	0.2156	−3.481	0.01
Condition	−0.5573	0.2611	−2.134	**0.05**

Table A.6: 500-600 ms window (ColAs).

	Coef β	SE(β)	z Value	p
Intercept	−0.5138	0.2358	−2.179	0.05
Condition	−0.5103	0.2530	−2.017	**0.05**

Table A.7: 600-700 ms window (ColAs).

	Coef β	SE(β)	z Value	p
Intercept	−0.3696	0.1989	−1.858	0.07
Condition	−0.3648	0.2687	−1.358	0.2

Table A.8: 700-800 ms window (ColAs).

	Coef β	SE(β)	z Value	p
Intercept	−0.11631	0.17305	−0.672	0.6
Condition	−0.09053	0.22241	−0.407	0.7

Table A.9: 800-900 ms window (ColAs).

	Coef β	SE(β)	z Value	p
Intercept	0.601579	0.243706	2.469	0.05
Condition	0.002608	0.194213	0.013	0.99

Table A.10: 900-1000 ms window (ColAs).

	Coef β	SE(β)	z Value	p
Intercept	1.521	0.277	5.476	0.001
Condition	0.027	0.232	0.116	0.91

Table A.11: 1000-1100 ms window (ColAs).

	Coef β	SE(β)	z Value	p
Intercept	2.646	0.3617	7.317	0.001
Condition	−0.118	0.3625	−0.326	0.8

Table A.12: 1100-1200 ms window (ColAs).

	Coef β	SE(β)	z Value	p
Intercept	−2.644	0.4005	−6.602	0.001
Condition	0.3861	0.4059	0.951	0.4

Table A.13: 0-100 ms window (RelAs).

	Coef β	SE(β)	z Value	p
Intercept	-2.4129	0.346	−6.973	0.001
Condition	0.2283	0.3154	0.724	0.5

Table A.14: 100-200 ms window (RelAs).

	Coef β	SE(β)	z Value	p
Intercept	−2.396	0.359	-6.662	0.001
Condition	−0.052	0.409	−0.129	0.9

Table A.15: 200-300 ms window (RelAs).

	Coef β	SE(β)	z Value	p
Intercept	−2.3481	0.4011	−5.854	0.001
Condition	−0.0906	0.4660	−0.194	0.9

Table A.16: 300-400 ms window (RelAs).

	Coef β	SE(β)	z Value	p
Intercept	−1.9157	0.308	−6.220	0.001
Condition	−0.2758	0.385	−0.715	0.5

Table A.17: 400-500 ms window (RelAs).

	Coef β	SE(β)	z Value	p
Intercept	−1.4899	0.2502	−5.955	0.001
Condition	−0.1653	0.307	−0.538	0.6

Table A.18: 500-600 ms window (RelAs).

	Coef β	SE(β)	z Value	p
Intercept	−1.1038	0.2117	−5.214	0.001
Condition	−0.2176	0.2394	−0.909	0.4

Table A.19: 600-700 ms window (RelAs).

	Coef β	SE(β)	z Value	p
Intercept	−0.6656	0.2243	−2.968	0.01
Condition	−0.2735	0.2764	−0.989	0.4

Table A.20: 700-800 ms window (RelAs).

	Coef β	SE(β)	z Value	p
Intercept	−0.2965	0.2363	−1.254	0.3
Condition	−0.3176	0.2781	−1.142	0.3

Table A.21: 800-900 ms window (RelAs).

	Coef β	SE(β)	z Value	p
Intercept	-0.002535	0.219697	−0.012	0.992
Condition	0.024452	0.255372	0.096	0.93

Table A.22: 900-1000 ms window (RelAs).

	Coef β	SE(β)	z Value	p
Intercept	0.4378	0.2108	2.077	0.05
Condition	0.270	0.2842	0.950	0.4

Table A.23: 1000-1100 ms window (RelAs).

	Coef β	SE(β)	z Value	p
Intercept	0.8339	0.2245	3.714	0.001
Condition	0.2938	0.3065	0.958	0.4

Table A.24: 1100-1200 ms window (RelAs).

	Coef β	SE(β)	z Value	p
Intercept	−2.3282	0.4181	−5.568	0.001
Condition	−1.0354	0.7545	−1.372	0.2

Table A.25: 0-100 ms window (MaxAAs).

	Coef β	SE(β)	z Value	p
Intercept	−2.2196	0.3658	−6.068	0.001
Condition	−0.9627	0.6905	−1.394	0.2

Table A.26: 100-200 ms window (MaxAAs).

	Coef β	SE(β)	z Value	p
Intercept	−2.4963	0.4669	−5.346	0.001
Condition	−0.1928	0.6307	−0.306	0.8

Table A.27: 200-300 ms window (MaxAAs).

	Coef β	SE(β)	z Value	p
Intercept	−2.73835	0.55665	−4.919	0.001
Condition	0.01886	0.70135	0.027	0.98

Table A.28: 300-400 ms window (MaxAAs).

	Coef β	SE(β)	z Value	p
Intercept	−2.6745	0.7237	−3.696	0.001
Condition	0.0869	0.8501	0.102	0.92

Table A.29: 400-500 ms window (MaxAAs).

	Coef β	SE(β)	z Value	p
Intercept	−2.4254	0.6117	−3.965	0.001
Condition	0.1436	0.6750	0.213	0.9

Table A.30: 500-600 ms window (MaxAAs).

	Coef β	SE(β)	z Value	p
Intercept	−1.7779	0.5283	−3.365	0.001
Condition	−0.4550	0.6239	−0.729	0.5

Table A.31: 600-700 ms window (MaxAAs).

	Coef β	SE(β)	z Value	p
Intercept	−0.6236	0.5369	−1.161	0.3
Condition	−0.9260	0.5516	−1.679	0.1

Table A.32: 700-800 ms window (MaxAAs).

	Coef β	SE(β)	z Value	p
Intercept	−0.1039	0.4572	−0.227	0.9
Condition	−0.8658	0.3600	−2.405	**0.02**

Table A.33: 800-900 ms window (MaxAAs).

	Coef β	SE(β)	z Value	p
Intercept	0.3869	0.3334	1.160	0.3
Condition	−0.6584	0.2355	−2.796	**0.006**

Table A.34: 900-1000 ms window (MaxAAs).

	Coef β	SE(β)	z Value	p
Intercept	1.0729	0.4492	2.389	0.02
Condition	−0.6121	0.3160	−1.937	0.06

Table A.35: 1000-1100 ms window (MaxAAs).

	Coef β	SE(β)	z Value	p
Intercept	1.7681	0.5790	3.054	0.001
Condition	−0.4269	0.4869	−0.877	0.4

Table A.36: 1100-1200 ms window (MaxAAs).

	Coef β	SE(β)	z Value	p
Intercept	−2.4408	0.3866	−6.314	0.001
Condition	0.3098	0.4716	0.657	0.6

Table A.37: 0-100 ms window (MinAAs).

	Coef β	SE(β)	z Value	p
Intercept	−2.5542	0.4416	−5.784	0.001
Condition	0.4086	0.5295	0.772	0.5

Table A.38: 100-200 ms window (MinAAs).

	Coef β	SE(β)	z Value	p
Intercept	−2.3699	0.4119	−5.754	0.001
Condition	0.2253	0.5302	0.425	0.7

Table A.39: 200-300 ms window (MinAAs).

	Coef β	SE(β)	z Value	p
Intercept	−2.2276	0.4144	−5.375	0.001
Condition	0.3952	0.537	0.736	0.5

Table A.40: 300-400 ms window (MinAAs).

	Coef β	SE(β)	z Value	p
Intercept	−1.85197	0.35004	−5.291	0.001
Condition	−0.09342	0.59706	−0.156	0.9

Table A.41: 400-500 ms window (MinAAs).

	Coef β	SE(β)	z Value	p
Intercept	−1.5010	0.3207	−4.681	0.001
Condition	−0.2248	0.5317	-0.423	0.7

Table A.42: 500-600 ms window (MinAAs).

	Coef β	SE(β)	z Value	p
Intercept	−1.2067	0.2570	−4.695	0.001
Condition	−0.2531	0.4836	−0.523	0.7

Table A.43: 600-700 ms window (MinAAs).

	Coef β	SE(β)	z Value	p
Intercept	−0.7741	0.2313	−3.346	0.001
Condition	−0.2821	0.5355	−0.527	0.6

Table A.44: 700-800 ms window (MinAAs).

	Coef β	SE(β)	z Value	p
Intercept	−0.388922	0.212360	−1.831	0.07
Condition	−0.004901	0.331717	−0.015	0.99

Table A.45: 800-900 ms window (MinAAs).

	Coef β	SE(β)	z Value	p
Intercept	−0.01706	0.22433	−0.076	0.94
Condition	0.04098	0.27272	0.150	0.9

Table A.46: 900-1000 ms window (MinAAs).

	Coef β	SE(β)	z Value	p
Intercept	0.3794	0.3003	1.263	0.3
Condition	0.2575	0.3704	0.695	0.5

Table A.47: 1000-1100 ms window (MinAAs).

	Coef β	SE(β)	z Value	p
Intercept	0.8525	0.2683	3.178	0.001
Condition	0.4100	0.4451	0.921	0.4

Table A.48: 1100-1200 ms window (MinAAs).